17th International Conference on Spatial Information Theory

COSIT 2026, September 22–25, 2026, York, UK

Edited by

Sabine Timpf
Gabriele Filomena
Armand Kapaj
Rui Zhu
Nicholas A. Giudice
Ed Manley

LIPIcs – Vol. 393 – COSIT 2026 www.dagstuhl.de/lipics

Editors

Sabine Timpf
Institute of Geography,
University of Augsburg, Germany
sabine.timpf@geo.uni-augsburg.de

Gabriele Filomena
Department of Geography and Planning,
University of Liverpool, UK
gabriele.filomena@liverpool.ac.uk

Armand Kapaj
Department of Geography,
University of Zurich, Switzerland
armand.kapaj@geo.uzh.ch

Rui Zhu
School of Geographical Sciences,
University of Bristol, UK
rui.zhu@bristol.ac.uk

Nicholas A. Giudice
Spatial Computing,
University of Maine, Orono, ME, USA
nicholas.giudice@maine.edu

Ed Manley
School of Geography,
University of Leeds, UK
e.j.manley@leeds.ac.uk

ACM Classification 2012
Computing methodologies → Model development and analysis; Computing methodologies → Intelligent agents; Computing methodologies → Spatial and physical reasoning; Computing methodologies → Ontology engineering; Computing methodologies → Mobile agents; Theory of computation → Theory and algorithms for application domains; Theory of computation → Data modeling; Information systems → Geographic information systems

ISBN 978-3-95977-438-3

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/978-3-95977-438-3.

Publication date
September, 2026

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists all publications of this volume in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://portal.dnb.de.

License
This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0): https://creativecommons.org/licenses/by/4.0/legalcode.
In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:
- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/LIPIcs.COSIT.2026.0
ISBN 978-3-95977-438-3 **ISSN 1868-8969** **https://www.dagstuhl.de/lipics**

LIPIcs – Leibniz International Proceedings in Informatics

LIPIcs is a series of high-quality conference proceedings across all fields in informatics. LIPIcs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

Editorial Board

- Patricia Bouyer-Decitre (Université Paris-Saclay, CNRS, ENS Paris-Saclay, FR)
- Roberto Di Cosmo (Inria and Université Paris Cité, FR)
- Faith Ellen (University of Toronto, CA)
- Holger Hermanns (Universität des Saarlandes, Saarbrücken, DE and Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, DE)
- Laura Kovács (TU Wien, AT)
- Daniel Král' (Leipzig University, DE and Max Planck Institute for Mathematics in the Sciences, Leipzig, DE – Chair)
- Sławomir Lasota (University of Warsaw, PL)
- Meena Mahajan (Institute of Mathematical Sciences, Chennai, IN)
- Chih-Hao Luke Ong (Nanyang Technological University, SG)
- Eva Rotenberg (IT University of Copenhagen, DK)
- Pierre Senellart (ENS, Université PSL, Paris, France)
- Alexandra Silva (Cornell University, Ithaca, US)

ISSN 1868-8969

https://www.dagstuhl.de/lipics

Contents

Preface
Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas A. Giudice, and Ed Manley 0:ix

Regular Papers

From Markers to Meaning: A Cognitive–Symbolic Model of Cultural Placemaking
Maria Carolina Quintana Noriega and Christophe Claramunt 1:1–1:18

When Are Geographic Quantities Semantically Compatible with Each Other?: Towards a Constructive Theory of Amounts, Magnitudes, and Proportions
Simon Scheider, Shuai Wang, and Eric Top 2:1–2:23

Qualitative Spatial Representation for Material Documents
Mark Bell and John G. Stell 3:1–3:19

RCC-8 as a Benchmark for Diagrammatic Reasoning in Multimodal Foundation Models
Robert E Blackwell and Anthony G Cohn 4:1–4:20

An Extended Qualitative Model for Reasoning About 3D Objects Using Depth and Different Perspectives
Martí Ejarque and Zoe Falomir 5:1–5:20

On Spatial Reasoning and Perspective Transformation in Language Models
Haotong Zhang and Ian Pratt-Hartmann 6:1–6:20

LMM Modality Transfer: A Pre-Requisite for Autonomous GIS Agents
Ivan Majic, Zexian Huang, Franziska Hübl, Krzysztof Janowicz, Meilin Shi, Mina Karimi, Zilong Liu, and Alexandra Fortacz-Lazan 7:1–7:20

Mapping Cycling Experience in Valence–Arousal Space: A Geometric Approach to Appraisal Heterogeneity
Ayda Grisiute, Johanna Wörle, Panagiotis Mavros, and Martin Raubal 8:1–8:20

Understanding Building Layouts from Minimal 360° Views: Effects of Viewpoint Overlap, Sequencing, and Landmarks
Ahmed Aly and Jakub Krukar 9:1–9:23

Much of Geospatial Web Search Is Beyond Traditional GIS
Ilya Ilyankou, Stefano Cavazzi, and James Haworth 10:1–10:20

Georeferencing Non-Gazetteered Place Names Using Biological Specimen Records
Aneesha Fernando, Surangika Ranathunga, Kristin Stock, Raj Prasanna, and Christopher B. Jones 11:1–11:22

Extending ISO-Space to Annotate Spatial Information in Route Descriptions
Samira Goudarzi Ali Akbari, Stephan Winter, and Martin Tomko 12:1–12:23

Mapinou: A Game to Assess the Skill of Using Pan-Scalar Maps
Justin Berli and Guillaume Touya 13:1–13:19

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas A. Giudice, and Ed Manley
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

An Operator Framework for Real-Time Analytics of Streamed Fields
JJ Paul and Silvia Nittel 14:1–14:22

Short Papers

Through and Through: Corpus-Based Semantic Analysis of a Spatial Preposition
Brandon Bennett and Maria M. Hedblom 15:1–15:7

From Trajectories to Relations: Interpreting Overtake Manoeuvres with Large Language Models and Qualitative Spatial Representations
Jana Verdoodt, Changbo Zhang, Lars De Sloover, Haosheng Huang, and Nico Van de Weghe 16:1–16:8

Limits of Spatial Reasoning Using LLMs
Varis Ithivatana and Brandon Bennett 17:1–17:8

A Cross-Linguistic Evaluation of Cardinal Direction Reasoning in LLMs
Panagiota Gyftou, Robert E. Blackwell, Manolis Koubarakis, and Anthony G. Cohn 18:1–18:9

Structuring the Underwater Environment: A Taxonomy of Aquatic Landmarks
Eva Nuhn and Sabine Timpf 19:1–19:8

GeoKG for GeoFM: Neuro-Symbolic Integration of Semantically-Enriched Geospatial-Temporal Data
Kai Moltzen, Debayan Banerjee, Julian Burmester, Anna Ehrenberg, Martin Kohler, Cedric Möller, Jann Pfeifer, Tilahun Abedissa Taffa, Hanna Marlene Schulz, Aida Usmanova, Patrick Westphal, and Ricardo Usbeck .. 20:1–20:9

Conceptual Vagueness in Human and LLM Representations: Evidence from Migration Distance Categorization
Songlin Wang, Krzysztof Janowicz, Ailin Benitez Cortes, Marion Borderon, Yingjing Huang, Mina Karimi, Zilong Liu, Annika Süß, and Patrick Sakdapolrak . 21:1–21:9

Habitual Use of Navigation Systems Impairs Spatial Learning Depending on Navigation Strategy
Qi Ying, Christopher Hilton, and Sara I. Fabrikant 22:1–22:8

Capturing Environmental Scale Spatial Cognition in Applied Settings: A Field Based Pilot with Firefighters
Chelsie McWhorter, Kathy Baylis, and Mary Hegarty 23:1–23:7

Context-Sensitive Spatial Deixis Through the Lens of Mental Models
Adwitiya Mandal, Zoe Falomir, and Kai-Florian Richter 24:1–24:8

Document Geocoding for Inferring Spatial Familiarity Within Reddit Text
Paddy Smith, Myles Gould, and Ed Manley 25:1–25:10

Learning a Map Under Positive Emotions: How Self-Reported Arousal and Individual Differences Relate to Route and Survey Knowledge
Veronica Muffato, Laura Miola, Francesca Pazzaglia, and Chiara Meneghetti 26:1–26:6

Representing Retail Wayfinding: A Requirement-Driven Study Design for Non-Visual Supermarket Navigation
Xiang Zhou and Sabine Timpf 27:1–27:7

Using Social Cues for Wayfinding: An Eye-Tracking Study of Attention Allocation and Evidence Accumulation in a Virtual Reality T-Intersection
Yu Qiao, Carlo Campagnoli, and Ed Manley 28:1–28:9

Feeling the Heat? Heat Stress Impairs Landmark but Not Route Knowledge in a VR Navigation Task
Alessandro J. Pierro, Sara I. Fabrikant, and Armand Kapaj 29:1–29:9

From Feature Attribution to Spatial Information Structure: Interpreting Urban Pedestrian Flows with Explainable AI
Atakilti Kiros, Achituv Cohen, Yuri Ribakov, and Israel Klein 30:1–30:8

Spatiotemporal Adaptive Learning for Flow-Based Graph Neural Networks
Fengjiao Li, Meiliu Wu, and Ana Basiri 31:1–31:9

The Same Street, Different Safety Judgments: Evaluative Framing Effects in Daytime and Nighttime Street View Imagery
Tong Qin, Nico Van de Weghe, Wim Fias, and Haosheng Huang 32:1–32:7

Decision-Making Under Uncertainty in Tsunami Evacuation: Individual Differences in Information Consultation and Risk Perception
Shota Iwama, Sihwan Lee, Hong Hu, and Takanori Nanahara 33:1–33:8

Simulating Path Integration with Continuous Attractor Neural Networks
Yulin Wu and Ed Manley 34:1–34:8

Towards a Validation Taxonomy for Agent-Based Models of Pedestrian Movement in Cities
Marcin Wozniak and Gabriele Filomena 35:1–35:10

From Sequential Decisions to Concurrent Processes: Advancing Human Navigation Models for Simulation
Sabine Timpf and Franziska Klügl 36:1–36:8

Preface

COSIT 2026 is the 17th Conference on Spatial Information Theory and was held from 22–25 September 2026 in York, United Kingdom. Established in 1993, COSIT is a biennial international conference series concerned with theoretical aspects of space and spatial information. It provides a forum for research on how space, time, spatial information, and temporal information are represented, reasoned about, communicated, experienced, and used by humans, agents, and computational systems.

COSIT takes place at a time when spatial information theory is increasingly asked to mediate between formal models, empirical studies of spatial cognition and behaviour, and rapidly developing computational systems. Current debates around GeoAI, geospatial knowledge graphs, embodied and agent-based spatial reasoning, foundation models and large language models, multisensory spatial cognition, and human–machine interaction all raise renewed theoretical questions about representation, interpretation, explanation, and situated spatial action.

Yet, spatial information also sits within a broader context. In response to pressing global challenges, researchers are being increasingly asked to consider and contribute to the understanding of people and space. Mapping our potential contributions – from spatial representation to production of spatial patterns – must become part of the practice of spatial information research. The theme for COSIT 2026, *"From Cognition to Flows"*, manifests this transition. It reflects on how the cognitive foundations of human behaviour hold an intrinsic role in shaping large-scale geographic flows and interactions, pertaining to complexity theories. Taken broadly, the theme challenges us to consider how representation impacts and interacts with the production of global phenomena.

As with tradition, COSIT 2026 is organized as an intensive single-track meeting. This format is central to the COSIT community: it encourages shared intellectual experience, sustained discussion across disciplinary boundaries, and the formation of new collaborations.

This year, the conference received 27 full paper submissions, a record number of 73 short paper submissions, and 9 submission for presenting journal papers. Each submission was reviewed by at least two members of the Programme Committee as well as the Programme Chairs. Where a submission was co-authored by a Programme Chair, the review and decision were handled by the remaining Chairs, and the conflicted member was excluded from all related discussions. Based on these reviews and subsequent discussions amongst the Programme Chairs, 14 full papers and 22 short papers were accepted for presentation and publication in this volume. An additional 19 submissions were accepted as posters, and 7 journal papers were accepted for presentation at the conference. Overall, the acceptance rate of submissions amounted to 36 %, with almost 48 % for full papers and 30 % for short papers. We also needed to desk-reject one submission due to improper use of AI and hallucinated references.

The papers in this volume demonstrate the breadth of contemporary spatial information theory. They address a variety of themes such as geospatial semantics, qualitative spatial reasoning, language and foundation models, spatial cognition, wayfinding and navigation, geographic information retrieval, spatial analysis and modelling, and spatial simulation.

As with the last COSIT, workshops and tutorials were incorporated as thematic sessions into the programme. We had four workshops during the conference on topics of responsible GeoAI, physical and perceived safety, spatial Human-Computer Interaction, and indoor

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas A. Giudice, and Ed Manley
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

spatial information. Alongside, we ran a Doctoral Mentorship Programme, encompassing 18 students in different stages of their PhD, with a special session featuring their research.

The programme is complemented by invited keynote talks from Professors Hugo Spiers, Laura Alessandretti, and Krzysztof Janowicz, addressing spatial cognition, mobility flows, and GeoAI, respectively. Together with workshops, the doctoral programme, and posters, these events broaden the scope of the conference and provide additional spaces for networking and discussion, especially for early-career researchers.

We thank all authors for submitting their work to COSIT 2026 and for contributing to the rigorous scientific quality of the conference. We are grateful to the Programme Committee and additional reviewers (see below) for their careful and constructive reviews. We also thank the local organizing team, the COSIT Steering Committee, the invited speakers, workshop and doctoral-colloquium organizers, sponsors, and Dagstuhl Publishing for their support in preparing this volume.

July 2026

Sabine Timpf
Gabriele Filomena
Armand Kapaj
Rui Zhu
Nicholas A. Giudice
Ed Manley

Organization

General Chair

Ed Manley

Local Organizing Committee

Gabriele Filomena
Lex Comber
Ed Manley

Programme Chairs

Sabine Timpf
Gabriele Filomena
Rui Zhu
Nicholas A. Giudice
Armand Kapaj

Themed Session Chair

Clio Andris

Doctoral Mentorship Chair

Crystal Bae

Communications Chair

Armand Kapaj

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas A. Giudice, and Ed Manley
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

List of Reviewers

Alexander Klippel
Alexis Comber
Amy Griffin
Andrew Crooks
Angela Schwering
Anthony G Cohn
Antoni Moore
Armand Kapaj
Arvid Horned
Bernd Ludwig
Carolynne Hultquist
Carsten Keßler
Chenyu Zuo
Christian Kray
Christophe Claramunt
Christopher Hilton
Clio Andris
Crystal Bae
Daniel R Montello
Ed Manley
Eliseo Clementini
Eva Nuhn
Franz-Benjamin Mocnik
Gabriele Filomena
Gamze Dane
Géraldine DEL Mondo
Gerhard Navratil
Gilberto Camara
Grant McKenzie
Guoray Cai
Henrikki Tenkanen
Heshan Du
Hoda Allahbakhshi
Ioannis Giannopoulos
Jakub Krukar
Jeremy G Morley
Johannes Scholz
John G. Stell
Judith Verstegen
Kai-Florian Richter
Kathleen Stewart
Krzysztof Janowicz
Leila De Floriani
Manolis Koubarakis
Maria M Hedblom
Markus Kattenbeck
Martin Raubal
Martin Tomko
Matt Duckham
May Yuan
Michela Bertolotto
Nicholas Giudice
Nico Van de Weghe
Peter Kiefer
Qi Ying
Reinhard Moratz
René Westerholt
Ron Mahabir
Ross Purves
Rui Li
Rui Zhu
Sabine Timpf

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas A. Giudice, and Ed Manley
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Sara Fabrikant

Sara Migliorini

Saskia Kuliga

Simon Scheider

Somayeh Dodge

Stefano Borgo

Stephan Winter

Stephen Law

Till Mossakowski

Toru Ishikawa

Tumasch Reichenbacher

Urska Demsar

Vanessa Brum-Bastos

Weiming Huang

Zhangyu Wang

Zoe Falomir

From Markers to Meaning: A Cognitive–Symbolic Model of Cultural Placemaking

Maria Carolina Quintana Noriega
School of Architecture, Art and Design, Tecnologico de Monterrey, Mexico City, Mexico

Christophe Claramunt
Arts et Métiers Institute of Technology, Paris, France
Naval Academy Research Institute, Brest, France

Abstract

This paper examines how cultural and artistic interventions contribute to the reappropriation of urban space by reorganizing attention, navigation, and the diffusion of place narratives. Comparing San Cristóbal de Las Casas (Mexico) and Le Panier (Marseille, France), we show that visible, meaning-rich markers act as attentional anchors, with micro-behaviours contributing to broader patterns of representation. Building on these empirical findings, we introduce a cognitive–symbolic model grounded in an anchor ontology that captures visibility, thematic resonance, persistence, and contextual relations, together with five indicators: Attentional Gain, Representational Reach, Neighbourhood Attentional Intensity, Counter-narrative Index, and Narrative Corridor that link micro-level salience to macro-level narrative prominence. Across both cities, the results reveal a dual regime in which curated, high-visibility anchors structure specific routes, while dispersed counter-narrative markers also persist. The model not only supports computable representations of place but also outlines pathways for integrating empirical data to better understand how cultural expressions shape the spatial narratives through which urban environments become meaningful.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Cultural Placemaking, Urban Street Art, Attentional ecology, Ontology

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.1

1 Introduction

Spatial information theory has undergone significant evolution over the past decades, expanding from its initial focus on geometric and topological representations of space to encompass cognitive, experiential, social and cultural dimensions. This shift reflects a broader epistemological transformation in the way urban environments are conceptualized, not merely as containers of physical infrastructure, but as dynamic arenas of meaning, memory, and social interaction [3, 9]. Central to this transformation is the notion of place, which contrasts with abstract conceptions of space. While space is often treated as a neutral, measurable entity, place is imbued with human experience, symbolic significance, and emotional resonance [19, 21].

In urban contexts, places are shaped not only by planning and architecture but also by the everyday practices, narratives, and cultural expressions of their inhabitants. Building on Jacobs' critique of rationalist urban planning and her emphasis on lived experience [8], we adopt a perspective where urban environments are shaped by everyday practices and symbolic expressions. Contemporary urban design approaches similarly foreground walkability, visibility, and human-scale experience [1]. These ideas have been foundational to contemporary urban design movements such as New Urbanism and Smart Growth, which prioritize walkability, mixed-use zoning, and human-scale development. This foregrounds the political and symbolic dimensions of public space, highlighting that urban environments are

© Maria Carolina Quintana Noriega and Christophe Claramunt;
licensed under Creative Commons License CC-BY 4.0

sites of citizenship and collective identity. In this paper, we use the term cultural placemaking to denote placemaking processes specifically driven by symbolic and artistic expressions in public space.

Yet, despite these advances, a conceptual and empirical gap persists: we know comparatively less about how micro-level markers of meaning and visibility, such as murals, stencils, inscriptions, and other cultural-artistic interventions, relate to macro-level patterns that are observable at the city scale. These patterns include not only pedestrian/tourist movement but, crucially for today's information-rich cities, flows of attention and representation (e.g., tour itineraries, photography and social-media traces) through which places are experienced and communicated.

This paper addresses that gap through a comparative inquiry into two culturally layered urban contexts: San Cristóbal de Las Casas, a city in Chiapas, Mexico, and Le Panier, a historic neighbourhood in Marseille, France, representing distinct yet comparable environments where artistic expression functions as a form of urban reappropriation. A central question is: How can symbolic and cultural dimensions of placemaking be represented within a semantic enrichment of urban space? To address this question, we investigate how micro-scale cues of visibility and meaning relate to broader patterns of attention and representation observable at the neighbourhood and city scale. We adopt an exploratory and interpretative stance, aiming to develop and exemplify a conceptual and modelling framework rather than to provide a validated empirical assessment. The intention is not to infer individual perceptual states or subjective interpretations of urban environments. Instead, street art is approached as a publicly available and materially situated trace that operates at the interface between perception and meaning. The focus is on how such markers contribute to shared attentional environments and representational structures within neighbourhoods, rather than on how they are individually interpreted or experienced. We conceptualize artistic interventions in public space as cognitive cues, salient, meaning-rich anchors that shape how urban environments are experienced, represented, and communicated by humans. By acting as attentional anchors, murals, stencils, and inscriptions modulate perceptual salience, prompting micro-behavioural effects (pauses, detours, photo taking) that aggregate into macro-level flows of representation and movement (e.g., inclusion in guided tours, higher photographic density, greater narrative visibility across media). This positions cultural placemaking as a cognition-flows mechanism: symbolic markers influence cognitive processing (selection, memory, interpretation) and thereby reconfigure informational and mobility trajectories within the city. These effects are approached through qualitative observation and representational traces rather than through direct behavioural measurement.

The contribution of this paper is threefold. First, through detailed empirical analyses of San Cristóbal de Las Casas and Le Panier, it shows how cultural and artistic interventions organise attention in urban spaces, revealing distinct yet comparable attentional ecologies shaped by visibility, thematic resonance, and location. Second, drawing directly from these grounded observations, we develop a cognitive–symbolic model of cultural placemaking, articulated through an anchor ontology that captures the key properties governing the salience of cultural markers, together with a set of indicators that describe how they participate in local attentional fields and broader representational flows. Third, by demonstrating how the model emerges from and returns to empirical material, the study shows how street-art expressions can be transformed into formal representations of place, enabling a bottom-up account of how attention, visibility, and narratives propagate through urban environments and become comparable across distinct urban contexts.

The remainder of the paper proceeds as follows. Section 2 reviews the shift from abstract spatial representations toward experiential and meaning-rich conceptions of place and frames placemaking as cultural reappropriation in urban space. Section 3 develops the two case studies whose empirical material grounds the subsequent conceptual developments. Section 4 builds directly on these cases to derive the cognitive–symbolic model, introducing the anchor ontology and the associated indicators that formalize how micro-scale cues of meaning contribute to broader patterns of narrative prominence. Section 5 concludes by discussing the implications for spatial information theory and urban research, and outlines avenues for advancing cognitive-based representations of urban places.

2 Cognitive Foundations and Cultural Reappropriation

This section builds on the humanistic-geographical tradition of place, in which place is understood not merely as a location, but as a meaningful relation between people, environment, and experience [19, 21]. It provides the theoretical and conceptual grounding for the study; the concepts discussed here motivate the analytical framework developed later, rather than serving as direct empirical claims about individual experience. While we draw conceptually on humanistic notions of place, the present study does not seek to model place as lived experience; rather, it focuses on symbolic and perceptually mediated elements that participate in processes of cultural placemaking. Recent work in GIScience and spatial information theory has sought to revisit and operationalize this tradition by examining how place meaning can be represented, compared, and reasoned about in formal and computational terms [25, 17]. Our contribution aligns with this line of inquiry by approaching cultural placemaking as a process through which symbolic and perceptual cues contribute to shared configurations of meaning, while remaining attentive to the interpretative and context-dependent nature of place.

Urban space has traditionally been represented through cartographic models centred on geometric accuracy, topology, and navigational utility, approaches that, while indispensable for analysis and planning, largely overlook the experiential and symbolic dimensions of urban life [9]. In contrast, work in cognitive and social geography has emphasized that people construct mental and socio-spatial representations of their environments: Lynch highlighted the role of cognitive maps [15], and it has been shown how such representations are shaped by everyday practices and social position, revealing how places are perceived, contested, and reappropriated [5]. Human geography further stresses that places emerge through familiarity, emotion, and lived experience [21], yet remain inherently relational and difficult to formalize [3, 4]. Across disciplines, research underscores the multifaceted nature of place [7], motivating ongoing efforts within spatial information theory to incorporate semantic approaches that account for meaning, memory, and cultural interpretation [9, 24]. Previous research models place semantics from narrative and textual corpora, highlighting representational traces as first-class evidence for place meaning [11]. This attention to representational sources resonates with long-standing findings in spatial cognition, where it has been shown how landmark, route, and survey knowledge are anchored in salient environmental cues [12], and how such cues act as cognitively privileged reference points that structure how routes are perceived, retained, and communicated [2].

Cultural placemaking is defined as a functional logic of appropriation where cultural, social, and political expressions transform urban spaces. Reappropriation of urban space, particularly in informal and marginalized contexts, has emerged as a practice of selective emphasis: communities foreground certain surfaces, corners, and routes to make them legible

as part of their environment [22]. Artistic expression plays a pivotal role in the reappropriation of urban spaces, serving as both a medium of resistance and a tool for identity reconstruction. Cultural interventions, such as murals, graffiti, and performative installations, transform overlooked or contested spaces into sites of memory, dialogue, and belonging. Heritage-based street art, for example, challenges dominant narratives and reclaims visibility for marginalized communities [16, 20].

Public art projects not only beautify urban landscapes but also promote inclusivity, civic engagement, and social justice [13, 14]. Informal cultural placemaking reflects the dynamic interplay between cultural production and spatial transformation, enabling communities to reclaim agency over their environments [6]. In this context, artistic expression becomes a critical lens through which urban space is reimagined, not as static infrastructure, but as a living canvas of cultural meaning and social interaction. Post-pandemic urban dynamics have amplified the relevance of feminist and intersectional narratives in public space. Inspired by movements such as MeToo and third-wave feminism [23], artistic interventions, particularly street art, have emerged as symbolic acts of resistance and identity formation. Placemaking becomes a multi-scalar phenomenon: it operates locally through acts of inscription and globally through circuits of representation, where visibility and thematic resonance determine how places circulate in touristic itineraries and social media.

This conceptual synthesis highlights placemaking as a process through which cultural expressions reorganize perception and meaning in urban space. By representing how visibility, symbolism, and spatial practices interact, we establish a foundation for analysing how these dynamics unfold in specific neighbourhoods. The following section turns to two empirical contexts where these mechanisms become legible.

3 Markers and Spatial Reappropriation in San Cristóbal and Le Panier

To explore how cultural placemaking recalibrates perception and circulates meaning, we examine two contrasting yet thematically resonant urban contexts: San Cristóbal de Las Casas and the historic neighbourhood of Le Panier in Marseille. The two study sites were selected deliberately for their contrasting yet analytically comparable roles as culturally expressive urban environments. San Cristóbal de Las Casas was selected for its strong Indigenous heritage and the presence of political, feminist, and religious expressions embedded in local street-art practices, while Le Panier was chosen as a historic neighbourhood in Marseille characterised by a multicultural urban fabric, heritage-oriented works, and a diversity of cultural expressions. Such representations are approached analytically through publicly observable cultural markers, rather than being elicited directly from inhabitants.

The empirical material for both case studies was collected through targeted, one-day field campaigns. In San Cristóbal de Las Casas, the first author, together with a small group of students, conducted a systematic survey in spring 2025, walking all streets within the selected central and surrounding boroughs to ensure full coverage of visible street-art landmarks. In Le Panier, the second author carried out an analogous one-day survey in winter 2026, following a preliminary visit conducted by the first author, again aiming at exhaustive coverage of the neighbourhood's walkable street network. This difference reflects logistical circumstances rather than a methodological distinction, and the same data collection and classification protocol was applied in both cases. In both cities, all identified items were photographed, geolocated, and then classified according to their most prominent medium (e.g. mural, stencil, graffiti, sticker, poster) and thematic category (e.g. feminist, political, cultural/heritage, religious, commercial) (almost 50 street-art landmarks for San Cristóbal de Las Casas, 47

for Le Panier in Marseille). The empirical scope of the survey was intentionally limited to publicly visible cultural markers that appeared to be deliberately placed to communicate symbolic, artistic, or narrative content in public space. This included murals, stencils, posters, graffiti, and similar inscriptions oriented toward a general audience. Highly repetitive or incidental markings (e.g., small stickers related to commercial branding or sports fandom) were not systematically recorded, unless they formed part of a coherent or recurring symbolic ensemble. Markers located on private surfaces were included when they were clearly directed at public visibility, while ephemeral or ambiguous cases were discussed jointly by the authors and resolved through qualitative judgment, guided by the analytical aims of the study. Each cultural marker was assigned a dominant thematic category, defined inductively from the empirical material, for analytical consistency across both sites, despite the acknowledged possibility of multiple symbolic attachments and interpretations. In addition to in-situ field observations, the study draws on a limited set of publicly accessible representational sources; such as online tour descriptions, cultural blogs, and image-sharing platforms, to assess how selected markers circulate beyond their immediate location. The qualitative indicators introduced in Section 4 were subsequently assigned jointly by the two authors, based on this shared inventory and visual documentation, in order to maintain consistency across the two cases while acknowledging the interpretive nature of the assessments.

San Cristóbal, a highland city with rich Indigenous heritage and a vibrant tradition of feminist and political activism, has become a laboratory for creative reappropriation, where educational institutions and community groups blend poetry, visual arts, and digital cartography to reclaim public space and articulate local narratives. In contrast, Le Panier, Marseille's oldest district, presents a dense, historically layered fabric marked by gentrification, tourism, and grassroots resistance. Its steep alleys and façades host street-art landmarks that oscillate between institutional beautification and subversive expression, creating a dual regime of visibility: curated works along tourist axes and oppositional messages in marginal alleys. By comparing San Cristóbal and Le Panier, we explore how distinct urban anatomies configure contrasting circuits of visibility and meaning, embedding local identity within broader urban imaginaries.

3.1 San Cristóbal de Las Casas

San Cristóbal de Las Casas, located in the highlands of Chiapas in southern Mexico, is a city renowned for its rich Indigenous heritage, colonial architecture, and vibrant traditions of political and artistic activism. Its unique cultural landscape makes it an experimental setting for exploring how artistic practices contribute to urban placemaking. Figure 1 situates San Cristóbal geographically, and the location of the urban area and its historical expansion up to the 19th century.

A starting hypothesis is that boroughs possess not only distinct identities but also specific territorial and attentional orientations (or vocations), which can be visualized through patterns of cultural expression. Mapping murals, graffiti, stencils, stickers, and posters reveals not just where art exists, but where attention is recruited and retained with feminist, political, and religious motifs drawing different publics into distinct circuits of noticing and pause. As an exploratory analytical lens, we examine whether certain areas exhibit stronger concentrations of activist, feminist, or resistance-oriented expressions, while others lean toward touristic, cultural, or decorative forms. Figure 2 illustrates representative examples of each category, highlighting their stylistic and material distinctions, while Figure 3 presents a series of cultural expressions of San Cristóbal de Las Casas, visualizing the different graphic categories and illustrating the diversity of street art.

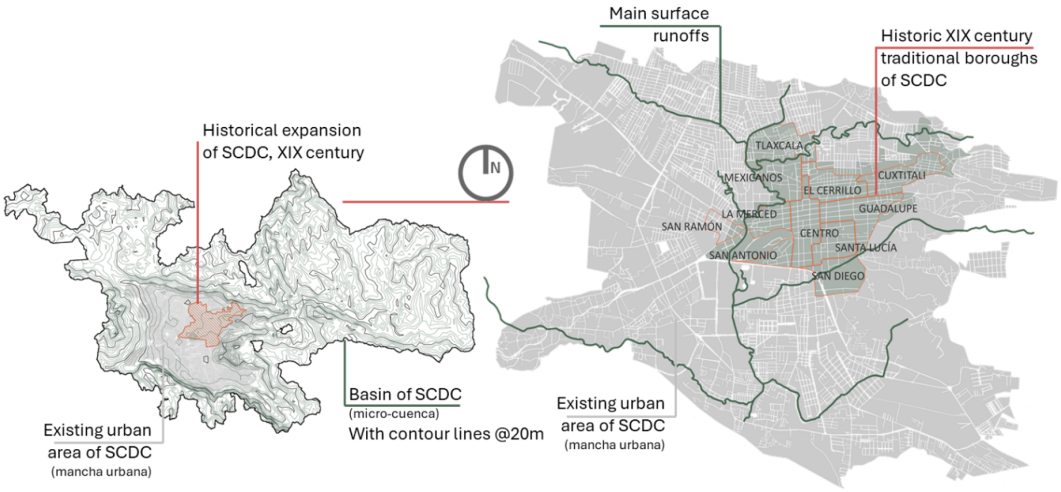


Figure 1 San Cristóbal de Las Casas.

Figure 2 Urban art categories in San Cristóbal de Las Casas.

The central historical area concentrates the high density of cultural murals alongside recurrent feminist and political messages (Historical centre and Barrio De Guadalupe), where visibility and footfall are greatest. The central boroughs operate as high-visibility attentional hubs in which large-format, legible imagery on façades and market fronts coexists with feminist stencils/posters, anti-Coca-Cola/water-rights critiques, and political iconography. While overtly feminist or politicized stencils are not the only forms present, they are consistently attested in the centre, and central murals still frequently display Mayan-inspired motifs, allegorical figures, and symbols of local belonging, feeding touristic circuits and public guides through repeated photographic capture and route inclusion [18]. The distribution is not spatially uniform: higher concentrations are observed in central areas characterized by high visibility.

In the northern neighbourhood of El Cerrillo (cluster to the top-left in Figure 3), a 17th-century historic area that acts as a threshold between the grid and the hills, resistance-oriented messages dominate the visual landscape, addressing autonomy, Indigenous memory, and social justice, often in continuity with political motifs observed in the centre. Feminist themes are also prominent, asserting gender equality and empowerment through bold imagery and slogans; religious-adjacent pieces appear where streets meet chapels or devotional routes, extending a sacred ambiance into the public realm. These expressions circulate through local networks and social media while remaining tightly embedded in neighbourhood life, producing resilient, memory-rich narratives [10]. This dynamic reflects a broader aspiration to reinvent the city according to collective desires, shaped by the practices of young people and artists.

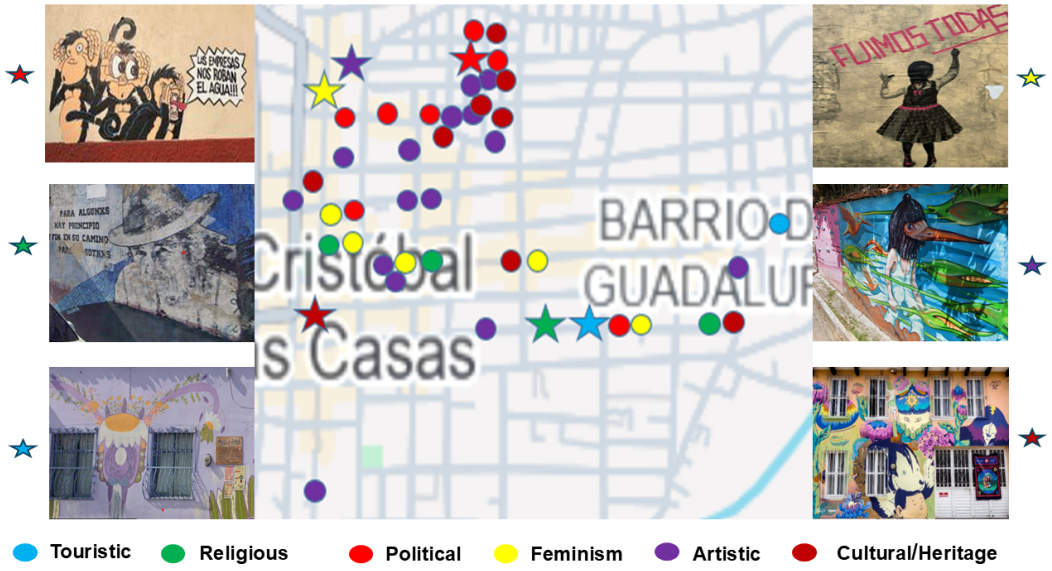


Figure 3 Examples of cultural art categories in San Cristóbal de Las Casas.

These findings indicate that urban art in San Cristóbal is not randomly dispersed but strategically positioned to reflect localized identities, contest dominant narratives, and reclaim space through visual storytelling. The data-driven baseline supports the hypothesis that media and themes correlate with borough vocation: central corridors combine cultural murals with recurrent feminist/political inscriptions; edge-of-centre neighbourhoods sustain resistance and devotional adjacencies; and peripheral sectors privilege cohesive community programs and long mural corridors. San Cristóbal de Las Casas does not exhibit a single canvas of cultural expressions, but a series of overlapping narratives that shape its strong social fabric.

3.2 Le Panier neighbourhood in the city of Marseille

The case of Marseille is centred on the historic district of Le Panier, a dense urban fabric characterized by steep alleys, layered architectural heritage, and vibrant street art (Figure 4). The quarter's morphology is a compact mesh of steep, narrow streets and tall buildings punctuated by places. The urban fabric retains fragments of medieval and early modern layers while accommodating contemporary commerce, galleries, and artisanal workshops.

Street art in Le Panier is remarkably diverse, encompassing murals, stencils, mosaics, posters, and ephemeral installations. Figure 5 illustrates the diversity of street art in Le Panier and showcases a range of themes, artistic and cultural heritage for the most prominent but also feminist, political, multicultural and touristic, reflecting the neighbourhood's layered identity and social dynamics. Framed as part of Marseille's "open-air museum", they blend decorative art with local heritage references to appeal to casual visitors and cruise-tourism audiences [10]. These works are often highlighted in artist-led, thematic tours focused on critical histories and alternative spatial narratives, which cater to niche publics and reinforce the alleys' mnemonic depth.

Le Panier exhibits a dual regime of attention: curated, large-format murals within the central belt and along tourist axes operate as institutional attention sources, repeatedly photographed and circulated in guides; while small-scale street-art landmarks serve as

Figure 4 Le Panier, Marseille.

attentional counter-sources, encoding feminist and anti-colonial narratives that attract niche publics and critical tours. The most visible works are concentrated around the streets extending north from Place de Lenche (from the feminist anchor example, Figure 5), and the vicinity of La Charité (cultural and artistic anchor examples, Figure 5), areas with high tourist traffic. Political, identity-driven messages and feminist anchors are often positioned close to the central areas of Le Panier, confirming that counter-messages are not displaced outward from the core. In the periphery of Le Panier, counts are lower, but surfaces are more cohesive (multi-artist or extended walls), yielding a distinct mode of legibility that contrasts with the centre's fine-grained density. This spatial pattern reflects a deliberate strategy: artists choose discreet locations to convey alternative narratives, creating a contrast with highly visible, tourist-oriented works. This duality between institutional visibility and grassroots subversion underscores the tension between urban branding and cultural authenticity [14]. However, this instrumentalization of street art by urban policies is accompanied by local resistance. This reveals a correlation between art forms (calligraffiti, stencil, collage) and territorial functions: decorative works tend to occupy tourist axes, while a few contestatory pieces are found on the margins of Le Panier [10].

The evolution of Le Panier, from a stigmatized zone to a tourist-facing "open-air museum", illustrates the complex dynamics of urban transformation. Gentrification has also led to the displacement of long-term residents and the rise of luxury boutiques, raising concerns about cultural authenticity. Community gardens and grassroots initiatives offer resistance, reclaiming space for participatory citizenship and social cohesion.

This duality reveals how artistic media correlate with territorial role orientations, echoing the typological mapping observed in San Cristóbal. These activities transform public space into a medium for storytelling and symbolic negotiation, where art becomes a tool for reclaiming visibility and contesting gentrification. The juxtaposition of curated tourist ambiances and lived community routines underscores the tension between urban branding and cultural authenticity. Artistic interventions, spatial storytelling, and forms of community

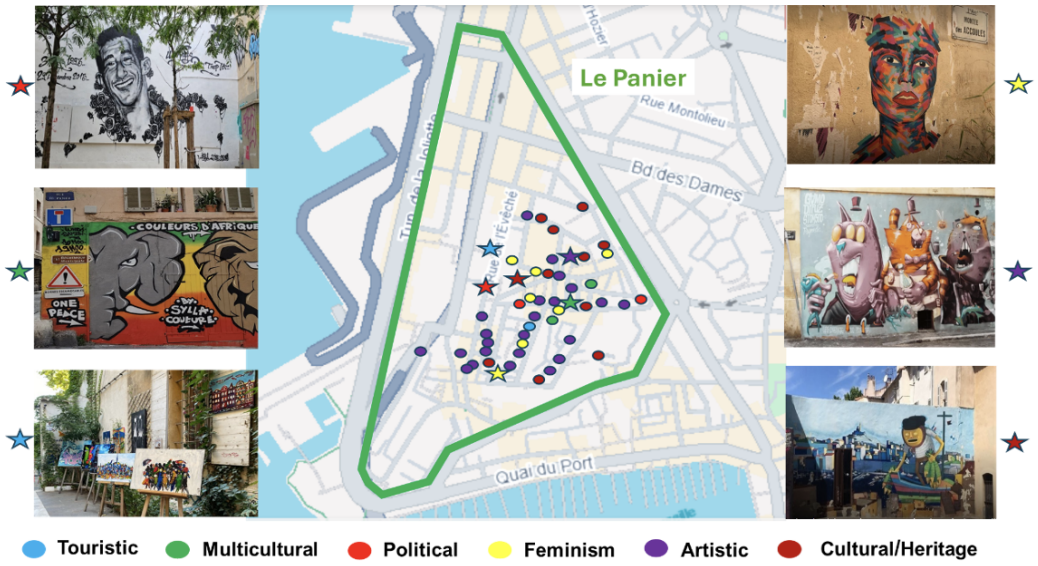


Figure 5 Urban art distribution and examples in Le Panier.

engagement can be described as contributing to configurations through which urban identities are expressed and sustained. The case underscores the need for inclusive policies that balance development with cultural continuity and support grassroots urbanism.

3.3 Discussion

The reappropriation of urban space through cultural expression in both San Cristóbal de Las Casas and Le Panier reveals deep social and territorial dynamics, and a distinctive logic. Many of the observed cultural markers appear informal and locally embedded, although the modes of production and governance of street-art practices vary and include both grassroots initiatives and institutionally supported interventions. The following discussion builds on qualitative patterns observed during fieldwork to interpret spatial configurations of cultural markers. In both cities, large, visually prominent murals situated along central or touristic axes serve as high-visibility anchors that stabilize dominant narrative heritage, identity, and civic symbolism. Conversely, peripheral or less travelled alleys host clusters of smaller, message-dense stencils, posters, and mosaics that sustain alternative or critical voices. These counter-narrative concentrations often mobilize themes such as feminism, resistance, and collective memory. In San Cristóbal, the coexistence of political, feminist, cultural, and artistic murals reflects a dense and activist-oriented social fabric in which Indigenous struggles, feminist mobilisations, and community-based initiatives collectively reclaim public space. In contrast, in Marseille's Le Panier, although feminist and multicultural expressions are also present and contribute to critical, often community-rooted voices, the visual field is more strongly structured by cultural-heritage and artistic motifs. This contrast highlights how differing social and cultural configurations in San Cristóbal and Le Panier are associated with distinct yet partly overlapping attentional logics and symbolic regimes. These thematic tendencies do not represent dominant categories, but they illustrate how cultural markers are embedded in, and expressive of, broader socio-territorial dynamics in each city.

These empirical patterns highlight a common attentional logic: visibility, thematic resonance, and spatial positioning interact to shape how cultural markers recruit perception and structure narrative pathways. Central anchors draw broad audiences through their

scale and accessibility, while dispersed micro-anchors attract attentive, often intentional engagement from more specific communities. Across both contexts, cultural markers thus operate not merely as decorative additions but as focal points that recode urban legibility, influencing pauses, detours, and the density of photographic or representational traces. Rather than claiming direct behavioural effects, our analysis proposes a conceptual link between micro-scale attentional salience at the level of individual markers and broader patterns of spatial organisation, such as recurrent narrative sequences. This interpretation is model-driven and exploratory rather than causal, and is intended as a conceptual basis for future work examining how such attentional configurations relate to practices in urban space. Accordingly, the patterns discussed here are not intended as generalizations of individual perception, but as ecological regularities emerging from the spatial distribution, visibility, and circulation of cultural markers in shared urban contexts.

Taken together, the findings point to the need for a formal account capable of capturing how micro-scale cues of meaning relate to broader representational and mobility patterns. The analysis does not aim to reconstruct artists' intentions or to validate meanings; instead, it focuses on how meaning-laden markers, once placed in public space, participate in shared attentional and representational dynamics. The recurrence of attentional hubs, counter-narrative clusters, and thematic concentrations suggests underlying regularities that warrant explicit modelling. Section 4 builds on these comparative insights by formalizing these dynamics through the anchor ontology and a set of indicators designed to operationalize how visibility, thematic content, and spatial relations contribute to urban attentional ecologies.

4 Toward a Computable Account of Cultural Placemaking

4.1 Ontology of cultural placemaking

Urban art and cultural placemaking practices foreground symbolic and perceptual dimensions of urban space by transforming otherwise unnoticed elements into focal points of attention and meaning. Rather than treating urban space solely as a geometric or functional entity, such practices highlight how visibility, placement, and symbolic content mediate how space is publicly encountered, attended to, and narratively articulated. We suggest an integration of cultural markers into a formal spatial ontology designed to support the semantic enrichment of urban space. The ontology introduced here is not intended to model place in the humanistic sense as lived experience or as a bond between individuals and the world. Instead, it provides a representational abstraction of culturally mediated symbolic and attentional configurations that participate in processes of cultural placemaking. It captures perceived, visible, and circulating meanings as they emerge through public exposure and spatial context, without attempting to encode authorial intent, individual interpretation, or local value systems directly.

The ontology is deliberately parsimonious and qualitative. It defines a minimal set of attributes and relations sufficient to represent how cultural markers function as attentional anchors within urban environments. Descriptive expressions used in the empirical sections (e.g. "symbolic corridors" or "heritage hubs") are therefore treated as interpretative summaries of spatial configurations produced by the ontology, not as additional modelling primitives. The assignment of thematic categories and indicator values was conducted jointly by the two authors as part of an exploratory framework-building exercise aimed at applying and stress-testing the model, rather than producing validated or inter-subjectively calibrated measurements. Accordingly, the resulting classifications should be understood as illustrative instantiations of the framework.

This leads to the following guiding challenge: How can symbolic and cultural dimensions of placemaking be represented within a semantic enrichment of urban space? Addressing this question requires an abstraction layer capable of extending spatial representation beyond geometry and topology to include symbolic, perceptual, and contextual attributes. To address this challenge, cultural markers are modelled as anchors, that is, spatially situated symbolic entities endowed with a set of qualitative attributes and relations.

- Anchor: A $\subset$ SymbolicMarker
- Attributes:
  - loc: geographic coordinates (point/segment), borough
  - vis: visibility class $\in$ {low, medium, high}
  - theme: category $\in$ (feminism, heritage, political, artistic, religious, commercial)
  - tech: tech $\in$ {mural, stencil, graffiti, poster, sticker}
  - persist: persistence (ephemeral, seasonal, semi-permanent, permanent)
- Contextual relations:
  - near(civic_space), near(tour_axis), adj(church/place)
  - cluster_with(Anchor) (co-located anchors forming a local attentional field)

Together, these elements define a lower bound of formal description: a minimal structure required to represent how cultural markers become salient, clustered, and narratively connected within urban space.

The thematic categories used in the ontology are intentionally general and analytically motivated. They are not intended as normative or exhaustive cultural classifications, but as abstract descriptors chosen to support cross-city comparability across heterogeneous urban contexts. While these categories inevitably reflect interpretative conventions, the ontology is explicitly extensible: additional or context-specific thematic dimensions can be introduced when applying the framework to particular settings without altering its core structure. The visibility attribute (vis) reflects the combined effect of physical salience (e.g., scale, contrast, façade placement) and situational visibility during navigation (e.g., exposure along pedestrian paths or frequently traversed streets). These aspects are considered jointly and qualitatively rather than as separable metrics, in keeping with the exploratory nature of the approach. Persistence (persist) is treated as a generic descriptive property rather than as a fully observed temporal variable. Its valuation relies on material form and limited secondary indications rather than on longitudinal observation. More detailed temporal modelling would require repeated surveys and is left for future work.

Finally, the ontology deliberately adopts parsimonious descriptors. Adding further attributes could increase specificity but would reduce comparability across cities. For example, although aesthetic qualities play an important role in cultural placemaking, they are not explicitly modelled here. This reflects an intentional focus on attentional and representational properties that are more readily comparable across contexts. Aesthetic valuation could nevertheless be introduced as an additional attribute in future extensions of the model, without altering its overall structure.

The ontology proposes a lower bound of formal description for cultural anchors: a minimal set of attributes required to reason about meaning propagation without loss of generality. Building on this ontology, each anchor (A) is characterised by a set of basic indicators that describe how anchors participate in the attentional organisation of a neighbourhood. The indicators are qualitative by design to favour ordinal distinctions over continuous metrics. This approach supports cross-context comparability without relying on locally variable quantitative data. Next, several semantic indicators are defined and rely only on attributes already assigned to anchors (cf. appendix for valuation examples).

Attentional Gain (AG). AG expresses how strongly an anchor is likely to draw attention based on the properties already defined in its description. Attentional Gain thus distinguishes visibility from meaning by considering visibility as a composite, qualitative property grounded in both physical prominence and navigational exposure. It sharpens the separation between physical visibility and semantic resonance. AG offers a compact interpretation of an anchor's attention-pulling capacity directly grounded in its visibility, theme, and contextual relations. Attentional Gain is assigned using qualitative categories (Low, Moderate, High).

$$AG(A) = f(vis(A), theme(A), context(A)).$$

In practice, AG was valued qualitatively using a simple process applied to the two case studies. High AG was assigned to anchors that combine high physical visibility with thematically salient content (e.g., feminist, political, or emblematic heritage motifs) in exposed locations along main pedestrian axes. Moderate AG was used for anchors with intermediate visibility or less resonant themes (e.g., commercial), and Low AG for small or visually marginal anchors in peripheral locations. These categories were applied jointly by the two authors on the basis of the photographic inventory and contextual notes. For example, a large cultural heritage mural in Le Panier (Figure 5, MP6 in the Appendix), set on a wide and highly visible façade, with a heritage-oriented theme along a busy tourist corridor, gives a high Attentional Gain, making it a prominent anchor that consolidates a heritage-focused narrative.

Representational Reach (RR). RR extends the description beyond local visibility and accounts for the degree to which an anchor is likely to appear in public narratives. It reframes urban art as part of an informational ecosystem, combining tour_presence(A), whether the anchor appears in guided itineraries and cultural paths, with media_presence(A), its occurrence in publicly accessible visual or textual traces (e.g., photographs, blogs, or online listings).

$$RR(A) = g(tour_presence(A), media_presence(A)).$$

RR was assessed by cross-checking the field inventory with a light survey of publicly accessible materials (online tour descriptions, blogs, and photo-sharing platforms) for each city, supported by targeted keyword-based searches and assisted browsing, including the use of large language models to identify and aggregate relevant sources. Anchors explicitly mentioned in several tours and appearing across multiple distinct online media sources were assigned High RR; those present in one or two sources were given Moderate RR; and anchors with little or no trace in these materials were valued as Low RR. For example, a colourful artistic mural on the public market façade in the centre of San Cristóbal (Figure 3, SC3 in the Appendix) is regularly included in walking tours and widely photographed by both visitors and residents. This repeated presence in guided itineraries and visual traces gives the mural high RR. At the same time, smaller and less curated inscriptions persist in everyday neighbourhood spaces, circulating primarily through local networks and activist practices rather than through formal visitor-oriented narratives.

Neighbourhood Attentional Intensity (NAI). NAI denotes how clusters of anchors co-produce dense attentional fields within defined areas of interest, such as highly circulated streets, locations around key places (e.g., squares), or entire neighbourhoods (e.g., Barrio del Cerrillo, Barrio de Guadalupe and Le Panier as a whole). This metric captures the collective

attentional density generated when locally salient anchors (high AG) also circulate through broader narratives (high RR). For a given area N, Neighbourhood Attentional Intensity is valued as:

$$NAI(N) = h(\{AG(A_i),\ RR(A_i)\}\ A_i \in AA)$$

where each anchor A_i within a street, block, or neighbourhood contributes through its Attentional Gain (AG) and Representational Reach (RR). h denotes a qualitative aggregation over all anchors within the area, where neighbourhood intensity increases with the number and co-presence of anchors exhibiting high AG and/or high RR rather than through numerical accumulation.

For a given area NN, NAI is valued qualitatively based on the distribution of AG and RR across its anchors: High NAI when multiple anchors show both high AG and high RR; Moderate NAI with mixed moderate values; Low NAI when most anchors have low AG and/or RR.

In the upper slopes of Le Panier near Place des Moulins (Figure 5, touristic MP2, political MP1, cultural-heritage MP6 in the Appendix), local streets combine several moderates to high visibility street-art anchors. Despite varying individual AG, their spatial co-presence and frequent inclusion in tours/photography yield high NAI, creating a dense symbolic field through cumulative interaction rather than a single dominant anchor. At the neighbourhood scale, Barrio del Cerrillo similarly exhibits high NAI from clustered resistance-oriented anchors along its devotional and activist circulation routes.

Counter-narrative Index (CI). Using the thematic categories assigned to anchors, CI highlights areas where alternative or critical voices are spatially concentrated within defined areas of interest, streets, locations around key places, or entire neighbourhoods. It combines $NAI(N)$ (attentional density) with $prop_opp(N)$ (proportion of anchors expressing oppositional themes such as feminism or political resistance):

$$CI(N) = q(NAI(N),\ prop_opp(N)).$$

CI is valued qualitatively: High CI when an area shows high NAI and many counter-narrative anchors; Moderate CI with moderate NAI and moderate proportions; Low CI when NAI is low and/or few counter-narrative anchors. For example, El Cerrillo (San Cristóbal) combines clusters of feminist stencils, Indigenous-rights graffiti, and political posters, representing several oppositional themes, while exhibiting high NAI. This yields high CI, marking El Cerrillo as a prominent counter-narrative hub. At street scale, similar clustering around the central Place de La Charité and Place des Pistoles in Le Panier also produce high CI from concentrated feminist anchors.

Together, AG, RR, NAI, and CI describe the Attentional Ecology of an area, that is, the configuration of anchors and the ways they influence what becomes noticeable, lingered upon local interpretation, photographed, or incorporated into local narratives. This ecology arises directly from the attributes and relations already encoded in the ontology, without requiring additional modelling layers. In Le Panier, the attentional ecology shows a mixture of high-visibility heritage anchors near La Vieille Charité and political micro-anchors in secondary alleys. In San Cristóbal, it alternates between cultural murals (high AG, high RR) and local political clusters (moderate AG, high CI).

Narrative Corridor. It is defined as a path through the street network along which a recognizable sequence of anchors. A Narrative Corridor formalizes how meaning and visibility propagate along walkable paths through sequences of anchors, bridging micro-scale salience

to meso-scale abstractions. A Narrative Corridor denotes a sequence of symbolic anchors made salient by meaning, thematic coherence, and representational attention rather than geometric form or path landmarks:

$$C = (A_1, e_{12}, A_2, e_{23}, ..., e_{(k-1)k}, A_k)$$

where $A_1, \ldots, A_k \in AA$ are successive anchors encountered in order along contiguous street segments. A path C is derived from a description that narrates or implies this ordered succession of locations and anchors. Once modelled, C can be summarized by path level descriptors such as:

$$RR(C) = (RR(A_1), RR(A_2), ..., RR(A_n)).$$

Corridors are extracted empirically from publicly available narrative descriptions (tour itineraries, blogs, guides) rather than constructed artificially. The process follows three steps: (1) Extract named streets/places from tour descriptions mentioning sequential street art or landmarks; (2) Match these to street segments containing ontology anchors; (3) Order anchors along the route to form $C = (A_1, ..., A_k)$.

Numerous available walking tours describe a classic street-art path: "Après avoir découvert le bas du Panier, on monte par la montée des Accoules, on flâne place de Lenche [...], puis [...] la rue des Pistoles et la vieille Charité". This yields the corridor MP-feminist (Place de Lenche) → MP6 (Montée des Accoules) → MP-central (La Vieille Charité) with uniform RR(C) = (High, High, High), illustrating a tourist axis (from the feminism to the central anchors in Figure 5, cf. Figure 6 right). In San Cristóbal de Las Casas free walking tour reviews note routes weaving from the centre through "street art in neighbourhood El Cerrillo" toward "Santo Domingo market" along "Real de Guadalupe". This produces SC-central → SC4/5/6 (El Cerrillo cluster) with RR(C) = (Moderate, Low, Moderate), showing rising attentional intensity through political/religious anchors (cf. Table 1).

These path descriptors reveal how micro-salience propagates: tourist corridors maintain high RR uniformity; resident/alternative corridors show thematic variation. Established corridors attract new anchors, creating path-dependent symbolic structures and semantics (touristic vs. local) within the same street fabric. Areas with higher NAI and anchors with higher RR are likely to anchor these corridors, linking cognitive effects to observable representational flows in tours, blogs, and media.

4.2 Analytical contribution and applicability

The anchor-oriented ontology and its associated indicators offers a structured representational framework for describing how symbolic, perceptual, and attentional dimensions are configured and related within urban space. Figure 6 articulates this framework as a four-level attentional mechanism, linking perception, local meaning patterns, narrative sequences, and city-scale representational flows. At the micro scale, anchors operate as perceptual attractors that can redirect navigation processes. This moves beyond the purely visual descriptors common in place ontologies by integrating perceptual, symbolic, and situational properties into a single qualitative index that denotes what is cognitively impactful. The meso-level concerns how multiple anchors jointly structure the attentional ecology of an area. Such ecological modelling formalizes how clusters of anchors co-produce local meaning fields such as dense symbolic corridors, resistance-oriented alleys, or heritage-oriented hubs, and why specific micro areas become influential in shaping neighbourhood identity. The last level organizes anchors into narrative corridors. While cognitive routes foreground legibility or navigational

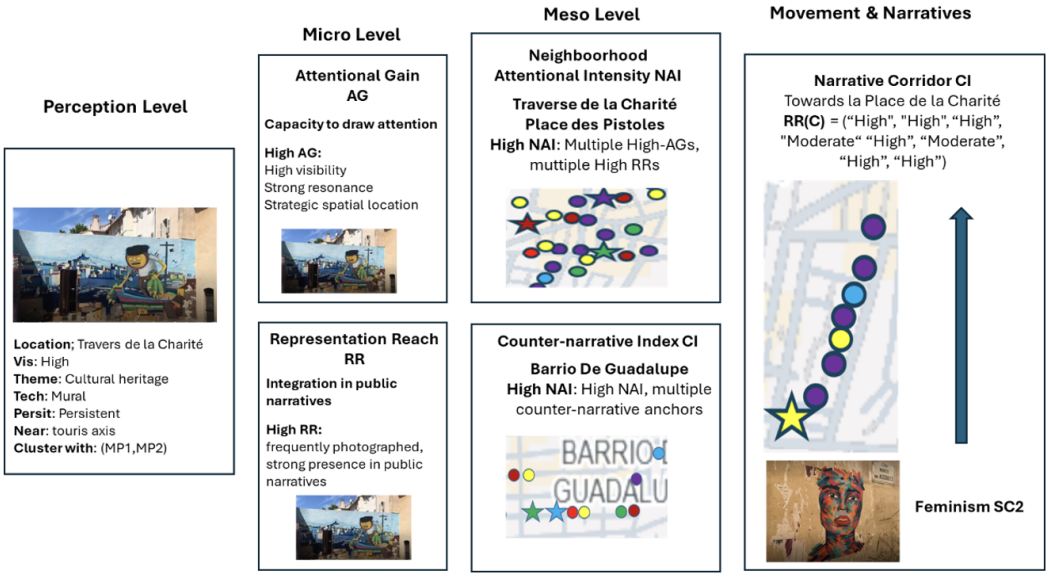


Figure 6 From perception to micro-macro levels, movements and narratives.

structure, these street-art narrative corridors foreground semantic and cultural properties. This makes it possible to reason about representational trajectories, how cultural meaning propagates along a route and to compare curated corridors with those that emerge organically.

Overall, the four-level structure provides a semantic and cognitive representation in which perceptual cues, local anchor configurations, and narrative pathways are related across scales within urban representational context. The proposed framework makes the following contributions:

- It integrates perception and semantics by treating meaning as an attention modifying property directly encoded in the ontology.
- It formalizes relationships across scales by representing how micro-scale cues are associated with meso-level configurations and broader representational patterns, rather than treating these levels as analytically independent.
- It introduces representational dynamics, allowing reasoning about how places become visible, remembered, or narratively prominent through cultural markers.

In doing so, the model expands the expressive capacity of spatial ontologies: not only where street-art things are, but how they are noticed, interpreted, sequenced, and circulated. Although derived from two empirical contexts, San Cristóbal and Le Panier, the ontology is intentionally abstract. Because all anchors share a common structure, attentional ecologies become structurally comparable across cities. Such comparisons move beyond descriptive analogies: they allow cities with distinct histories, morphologies, or artistic grammars to be examined through a shared symbolic vocabulary. This opens pathways for identifying recurrent spatial logics across cities: whether feminist inscriptions systematically gravitate toward transitional areas, whether heritage-oriented murals consistently anchor touristic corridors, or whether counter-narrative clusters tend to form in morphologically marginal pockets.

5 Conclusion

This paper examined how artistic and cultural expressions reorganize perception, attention, and meaning in urban space through a comparative study of San Cristóbal de Las Casas and Le Panier. The study illustrates how cultural markers can be represented as attentional anchors that participate in what becomes noticeable, memorable, and narratively salient within the city. Building on these empirical observations, we introduced a cognitive–symbolic model of cultural placemaking that formalizes how micro-scale cues of visibility and meaning aggregate into broader spatial patterns. The anchor ontology and its indicators provide a unified way for describing attentional ecologies within urban environments. It contributes a representational perspective in which cultural place is approached not merely as a physical location but as a dynamic configuration of expressive relations articulated through cultural markers. Together with narrative corridors and the four-level attentional mechanisms, the model links perceptual salience, local clustering, ordered sequences, and macro-level representational flows. Theoretically, this work contributes to spatial information theory by advancing a street-art model that integrates cognitive, symbolic, and narrative dimensions of place. Beyond aesthetics, an attention-based view of placemaking suggests new metrics, captured by the indicators developed here, that supports a semantic representation of urban cultural meaning. Cultural markers participate in the symbolic and attentional infrastructure of the city, influencing what is foregrounded in shared narratives and representations.

In both cities, despite their divergent socio-cultural and geopolitical contexts, common trends emerge: the emergence of feminist and heritage-based themes, the strategic placement of art in contested or transitional zones, and the tension between grassroots expression and institutional visibility. We observed recurrent spatial configurations: central anchors with high visibility and thematic prominence structure widely shared narratives. Peripheral or marginal clusters form dense counter-narrative fields that sustain alternative or resistant expressions. Narrative corridors, in turn, channel meaning along walkable paths.

The model still encompasses several limitations to address. First, the indicators are based on qualitative expert assessment rather than quantitative measurement of visibility, flows, or media propagation. Their value lies in conceptual clarity, but future empirical validation is needed. Second, the model abstracts from the temporal volatility of cultural expressions: markers appear, fade, are erased, or replaced, and this dynamism is only partially represented in our current framework. Through these developments, the model can contribute to a representation of cultural urban places not simply as spatial containers, but as dynamic, meaning-laden environments shaped by cultural expressions, narrative practices, and the collective reappropriation of space. By formalizing the links between perception, meaning, and representation, the model contributes a semantic and cultural-based representation of urban spaces.

Ethical note. All street-art markers analysed are publicly displayed works created for open visibility in urban space. The documentation process involved no interaction with individuals, no collection of personal data, and no recording of sensitive or private contexts. Our analysis operates at the level of aggregated spatial patterns and symbolic categories, ensuring that actors or communities are not identifiable. As such, the study does not raise significant ethical risks and adheres to standard principles of respectful and non-intrusive research in public environments.

References

1 Jordi Borja and Zaida Muxí. *El Espacio Público: Ciudad y Ciudadanía*. Electa, 2003.

2 David Caduff and Sabine Timpf. On the assessment of landmark salience for human navigation. *Cognitive Processing*, 9(4):249–267, 2005.

3 Helen Couclelis. People manipulate objects (but cultivate fields). In *Spatial Information Theory*, pages 65–77, 1992.

4 Tim Cresswell. *Place: A Short Introduction*. Blackwell, 2004.

5 Pedro Dias and Thibaud Ramadier. Representational structures as stances. *International Review of Social Psychology*, 31(1):24–45, 2018.

6 Magnus Fransberg, Rui Simões, and Paulo Raposo. Informal placemaking. *Journal of Urbanism*, 17(2):165–168, 2023.

7 Howard Frumkin. Healthy places. *American Journal of Public Health*, 93(9):1451–1456, 2003.

8 Carmen de J. Gómez Abarca. *Subversión Visual: Graffiti, arte urbano, gráfica politica y muralismo en San Cristóbal de las Casas*. Universidad de Ciencias y Artes de Chiapas, 2024.

9 Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

10 Alice Grondeau, Florence Pondaven, and Matthieu Boulpicante. Le street art à Marseille: entre communication urbaine, institutionnalisation et instrumentation. *Méditerranée*, 14456, 2023. URL: `https://journals.openedition.org/mediterranee/14456`.

11 Emad Haris, Anthony G. Cohn, and John Stell. Semantic perspectives on the Lake District writing: spatia ontology modelling and relation extraction for deeper insights. In *Proceedings of COSIT 2024*, volume 315 of *LIPIcs*, pages 11:1–11:19. Schloss Dagstuhl, 2024. `doi:10.4230/LIPIcs.COSIT.2024.11`.

12 Toru Ishikawa and Daniel R. Montello. Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of learned places. *Cognitive Psychology*, 52(2):93–129, 2008.

13 Min Kang. Street art for urban regeneration: cultivating community, identity, social and economic growth. *Journal of Student Research*, 14(1), 2025.

14 Soo Jin Lee and Liza Shulyakovskaya. Placemaking through artistic pedagogy: advancing anti-oppression in public spaces. In *Proceedings of the International Conference of the Learning Sciences (ICLS)*, pages 1644–1648, 2025.

15 Kevin Lynch. *The Image of the City*. MIT Press, 1960.

16 Nancy Macdonald. *Urban Governance and Street Art*. Routledge, 2016.

17 Franz-Benjamin Močnik. On the representation of places. *GeoJournal*, 88(4):4109–4126, 2023.

18 María de Lourdes Morales Vargas. Relatos a la espera: muralismo urbano en los espacios publícos de san cristóbal de las casas, chiapas. *LiminaR*, 18(1):61–81, 2020.

19 Edward Relph. *Place and Placelessness*. Pion, 1976.

20 Rafael Schacter. *The World Atlas of Street Art and Graffiti*. Yale University Press, 2014.

21 Yi-Fu Tuan. *Space and Place: the perspective of experience*. University of Minnesota Press, 1977.

22 Martijn Van Oostrum. Appropriating public space: transformation of public life and loose parts in urban villages. *Journal of Urbanism*, 15:84–105, 2022.

23 Nuria Varela. El tsunami feminista. *Nueva Sociedad*, 286:93–106, 2020.

24 Stephan Winter and Christian Freksa. Approaching the notion of place by contrast. *Journal of Spatial Information Science*, 5:31–50, 2012. `doi:10.5311/JOSIS.2012.5.90`.

25 Bo Zhao. Humanistic GIS: Toward a research agenda. *Annals of the American Association of Geographers*, 112(6):1576–1592, 2022.

A Appendix

The following tables illustrate the indicators applied to representative anchors discussed earlier (Figure 3 for San Cristóbal; Figure 5 for Le Panier). Anchor attributes (visibility, theme, contextual relations, persistence) were jointly assessed by the two authors based on field observations, in-situ positioning, and systematic photographic documentation. Attentional Gain (AG) reflects the authors' interpretation of perceptual and semantic prominence as observed during fieldwork, taking into account scale, clarity, and thematic distinctiveness. Representational Reach (RR) was established using a light survey of its presence across publicly accessible online materials, including walking-tour descriptions, cultural blogs, user-generated photo platforms, and image-rich search results. This provided a simple but consistent approximation of whether an anchor circulates beyond its immediate location and enters broader representational flows. Neighbourhood Attentional Intensity (NAI) and Counter-Narrative Index (CI) were derived from the distribution and thematic attributes of anchors within each area, providing a qualitative description of how visibility and thematic concentration are configured locally. Table 1 summarizes anchor values while Table 2 evaluates neighbourhood values for Barrio del Cerrillo and Centro for San Cristóbal de las Casas, and Le Panier in Marseille.

Table 1 Anchor-level indicators.

Anchor	City	Location	Media	Theme	AG	RR
SC1	San Cristóbal	Centro	Muralism	Heritage	High	High
SC3	San Cristóbal	Barrio del Cerrillo	Muralism	Artistic	High	High
SC2	San Cristóbal	Centro	Stencil	Feminism	Moderate	High
SC4	San Cristóbal	Barrio del Cerrillo	Graffiti	Commercial	Low	Low
SC5	San Cristóbal	Barrio del Cerrillo	Muralism	Political	High	Moderate
SC6	San Cristóbal	Barrio del Cerrillo	Muralism	Religious	Moderate	Low
MP1	Marseille	Le Panier	Muralism	Political	High	Moderate
MP3	Marseille	Le Panier	Mosaic	Multicultural	High	High
MP2	Marseille	Le Panier	Poster	Touristic	Low	Low
MP4	Marseille	Le Panier	Poster	Feminism	Moderate	Moderate
MP5	Marseille	Le Panier	Muralism	Artistic	High	High
MP6	Marseille	Le Panier	Muralism	Cultural heritage	High	High

Table 2 Place and neighbourhood-level indicators (NAI and CI).

City	Area	Anchors	NAI	Counter-narrative anchors CI
San Cristóbal	Barrio de Guadalupe	10	High	Moderate
San Cristóbal	Barrio del Cerrillo	26	High	High
Marseille	Place de la Charité & Place des Pistoles	22	High	High

When Are Geographic Quantities Semantically Compatible with Each Other?

Towards a Constructive Theory of Amounts, Magnitudes, and Proportions

Simon Scheider[1]
Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Shuai Wang
Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Eric Top
Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Abstract

Geographic quantities are more than just numbers. A particular semantics is required to interpret them, which is best captured by constructive types. From a practical viewpoint, numerical operations are only applicable to certain types of geographic quantities, while for others, they become meaningless, even if they are numerically well-defined. For example, GIS experts know that it is meaningless to multiply some quantities – measured, for instance, in people or square kilometers – with one another. In contrast, areas can meaningfully be multiplied by population densities to obtain the number of people living in that area. We currently lack a semantic theory that allows us to determine whether quantities represented in geodata models are compatible with one another for deriving new quantities by common GIS operations. In this paper, we distinguish three distinct types of quantities that are often collapsed into numbers: amounts, magnitudes, and proportions. For each type, we introduce a constructive first-order logic theory. We then define quantifiable experiments represented by different data sources drawn from a geodata repository, and apply the theory to determine the compatibility of those sources for deriving new quantities.

2012 ACM Subject Classification Computing methodologies → Reasoning about belief and knowledge

Keywords and phrases geographic quantities, semantics of measurement, GIS, geodata repositories

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.2

Supplementary Material *Dataset (Prover9 script and ArcGIS Pro project and workflow)*: `https://zenodo.org/records/20753038`

Funding Our research was funded by the European Research Council (ERC)'s HORIZON.1.1 program, grant id: 101170816.

Acknowledgements We would like thank our anonymous reviewers for their suggestions, which considerably improved the manuscript.

Declaration on GenAI/LLM use The authors used ChatGPT and Writefull on Overleaf to check grammar and to improve paraphrasing of manually written text. In no case were the ideas of this paper generated by AI.

1 Introduction

One of the central methodological gaps in geographic modeling and cartographic representation is the absence of explicit *validity criteria* for models of geographic quantities. Data-driven numerical practices alone do not provide explicit justifications for why models of those practices should be considered valid [38]. While GIS users know from training

[1] Corresponding author.

© Simon Scheider, Shuai Wang, and Eric Top;
licensed under Creative Commons License CC-BY 4.0

and experience that numerical operations on geographic data are only valid under certain conditions, the rules governing the selection, combination, and transformation of data are rarely made explicit. Consider, for example, the assessment of solar potential and current renewable energy use across Amsterdam's neighborhoods. To arrive at valid conclusions, it is necessary to account for the fact that mapped quantities differ in semantic type, spatial unit, and level of measurement, and are distributed across multiple geodata sources: neighborhood area, roof area suitable for solar panels, the number of existing rooftop panels, and household energy consumption.[2]

Closing this methodological gap has become increasingly important because GIS lacks general validity criteria [40] as well as widely applicable standards for interdisciplinary use [24]. The problem is further amplified by the growing reliance on data-driven GeoAI [21], where explicit information practices are increasingly replaced by self-supervised machine learning models, including geo-foundational models [22]. These models are expected to support a wide range of downstream purposes, yet it remains unclear *why* the practices they model support valid conclusions [40, 38].

Addressing this issue requires a theory of the semantics of quantities [40, 38]. Such a theory must answer practical questions such as (in our example): How can heterogeneous information sources be combined to model solar potential and its exploitation? Does it make sense to sum, subtract, multiply, or divide areal measurements, energy capacities, and panel numbers? If so, what kinds of quantities and spatial units result? Whether a workflow yields valid conclusions depends on whether its numerical inferences are valid, which in turn depends on the semantic compatibility of the quantities involved. Without such a theory, we risk producing quantifications that may be automated but are not justifiable. At a minimum, it must distinguish between valid and invalid numerical inferences, and between compatible and incompatible data sources in the context of a question [40].

Existing theories remain insufficient because they address only parts of this problem. We know, for example, that extensive quantities may be summed and represented in graduated symbol maps, whereas intensive quantities should not [37]. Yet current theory still amounts largely to an extended list of measurement levels [6]. Measurement theory [32] has a long tradition and has strongly influenced GIScience [6], while foundational ontologies of quantities have been proposed more recently [12, 36]. In parallel, fuzzy set theory laid the foundations for Qualitative Reasoning (QR), which seeks to automate qualitative reasoning about intervals on linear scales [50]. However, these approaches are neither sufficiently specific to geographic information nor do they provide criteria for assessing the compatibility of geographic quantities. In particular, they typically require committing to formalizations of quantities on the real number line [50]. Yet we know that extensive geographic quantities represent different mathematical structures, including experimental controls over mathematical lattices, cf. [48, 49]. Inspired by [43], we recently proposed a corresponding experimental model of geographic information [41], which also accounts for model validity [40]. However, we are still lacking criteria for assessing the semantic compatibility of quantities, leaving a central prerequisite for model validity unresolved.

To address this gap, this paper investigates the following research question: *Under what conditions can geographic quantities represented in geodata sources be combined to derive new valid quantities?* We answer this question by proposing a *constructive theory* of geographic quantities that specifies their *typology* (types of geographic quantities and their experimental structures), *formalization* (reconstructions constraining permissible operations and valid outcomes), and *compatibility* (criteria for determining valid combinations of geodata

[2] Such data is e.g. available at `https://maps.amsterdam.nl/`

sources). While our theory targets geographic quantity construction based on an experimental interpretation of geographic information concepts, it may have broader applicability beyond geography and GIS. To demonstrate its use for geographic quantitative reasoning, we revisit the Amsterdam example to determine the solar energy potential of roofs in the city center.

2 Preliminaries

This paper extends prior work on extensive quantities [48] and experimental measurement [41]. Because terms like "mass" [35], "quantity" [13], and "measure" [43, 7] are used ambiguously in geography and GIScience, we clarify our terminology first. Our approach diverges from conventions in physics and standard representational measurement theory [8, 32], which treat quantities as mappings from the world to measurement scales. Instead, it aligns with pluralist accounts of mereological quantities[3] [1, 12] and models them based on a constructivist epistemology[4].

2.1 Quantities

We consider a *quantity* to be an entity that can be used within a *controlled experiment* and that can be *quantified*.[5] For example, in a census (see examples in Table 1), we may control spatial regions and register their inhabitants. As this example shows, quantities need not be numbers or lie on a linear scale. The region of Amsterdam, its area, its population, and its population density are all different kinds of quantities.

A *quantity domain* is the set of all possible quantities of the same type that can be used in an experiment. Quantity domains are constructed bottom-up from measurement practices [18]. For example, *linear measurement scales*, such as the length (m) or area (m^2) scale, are special cases of quantity domains. We call their quantities *magnitudes*, and distinguish them from domains constructed by ratios of magnitudes (*proportions*, such as densities). Both magnitudes and proportions are *linear quantities* because they can be linearly ordered.

Quantities whose domains form a non-linearly ordered part–whole structure that is closed under sums, subtractions, intersections, and unions are called *amounts* (domain symbol A). Examples include amounts of material, mass, water, soil, energy, people, space, and time. For example, we can sum amounts of people to obtain larger ones, or intersect them to determine which individuals they have in common. Similarly, in a GIS, amounts of space (called regions) can be summed or intersected to generate new amounts of space, e.g., using spatial overlay [7]. In other words, amounts are *homeomerous* concepts [13]: parts, sums, and intersections of amounts are themselves amounts. This fundamentally distinguishes them from *objects*, such as buildings or social groups: if we divide a building in half, we do not obtain another building, and likewise, dividing a social group does not yield another group. In both cases, however, we do obtain the underlying amount of material or people

[3] The idea that multiple types of parthood – and hence multiple types of quantity – may coincide. For example, a forest can be conceptualized as an object as well as a quantity of trees, which can be measured by area, number of trees, and biomass, depending on which aspect of "forest" is relevant.

[4] More precisely, we use prior work on protophysics [18] and constructive logic [29, 31, 28] developed in the Erlangen school of methodical constructivism.

[5] This reflects ordinary linguistic usage of quantities (count and mass nouns) more closely [46] and, more importantly, makes explicit the mereological and constructive origins of quantities in geographic information [48]. In doing so, we draw on constructivist proposals by Janich [18], Gruninger's foundational ontology of units and measures [12], and key insights from Simons [42].

Table 1 Comparing amount, magnitude, and proportion with examples.[6]

	Amount (A)	**Linear quantity**	
		Magnitude (M)	**Proportion (P)**
construction	a mereological quantity whose domain forms a boolean lattice.	a linearly ordered quantity constructed from some amount.	a linearly ordered quantity constructed from pairs of magnitudes.
operations	can be summed, subtracted, intersected, and partially ordered by parts.	can be linearly ordered, summed, and subtracted, but does not form products or ratios.	can be linearly ordered, summed, subtracted, and form products and ratios.
properties	boolean lattice	linear (total) order	
formerly [48]	amount	archimedean magnitude	proportional magnitude
example	the region or population of a neighborhood in Amsterdam.	the size of the region or the number of people of a neighborhood in Amsterdam.	population density: the proportion of population and region size (area) of a neighborhood in Amsterdam.

that constitute them. Formally (see Appendix) [48], amount domains are partially ordered by *parthood* $\subseteq$, satisfying reflexivity, antisymmetry, and transitivity. They support *sums* $(+)$ and *intersections* $(*)$, forming a complete *Boolean lattice* [1]. *Overlapping* amounts $O(a_1, a_2)$ have non-empty intersections (see Appendix). Two distinguished elements are important: the empty amount 0^A ("nothing"), which is neutral for sums and absorbing for intersections, and the total amount 1^A ("everything"), which is neutral for intersections and absorbing for sums. Complements $(-a)$ are what is left when removing a from everything. *Subtraction* $(\backslash)$ is defined via complement and intersection. Importantly, amount lattices are non-linear (i.e., non-overlapping amounts exist). The resulting mereology is extensional and satisfies the supplementation principle [5]. More details can be found in the Appendix.

Magnitudes (domain symbol M), in contrast, allow us to measure the sizes of amounts and to compare them even when the corresponding amounts cannot be directly compared [48]. For this reason, magnitudes behave fundamentally differently from amounts. Because sums of amounts may overlap, their magnitudes do not necessarily sum in the same way as the amounts themselves. For example, summing a neighborhood region in Amsterdam with the larger region of Amsterdam that contains it yields the region of Amsterdam itself, but summing their magnitudes (areas) produces a magnitude larger than the area of Amsterdam. Furthermore, magnitudes do not form products (intersections) or ratios that yield new magnitudes. While intersecting regions is a basic technique of GIS, multiplying their sizes to generate new sizes is nonsense [42]. However, multiplying an area magnitude with a *proportion* (domain P) such as population density can be very useful.

Table 1 compares the terms *amount*, *magnitude*, and *proportion* as used in this paper, highlighting differences in their construction, admissible operations, and formal properties. The underlying idea originates in [48], though we simplified the terminology[7]. Magnitudes must also be distinguished from their *quantifications*. Quantification requires *measurement units*, which are specific kinds of *linear reference systems* [7]. For example, the gram specifies

[6] The table was adapted from [48] with the permission of the authors.
[7] We use magnitude instead of "Archimedean magnitude" and proportion instead of "Proportional magnitude".

a linear reference system for mass magnitudes [8]. It can be used to assign a number to a magnitude relative to the chosen unit. Thus *lb* and *g* correspond to different units that refer to the same magnitude domain (mass) [6]. Magnitudes are abstractions that are independent of any scale, representing a quantity under any chosen unit.

2.2 Experimental measurement of geographic quantities

Understanding spatial information requires more than examining a quantity of interest; it also requires analyzing its experimental structure, in particular what is *controlled*, what is *fixed* (held constant), and what is *measured* [43]. As suggested in [41], we thereby control, fix, or measure situations in an underlying experiment [26]. For example, interpreting data from a *tide gauge*, which *measures water level heights* in a time series, requires more than knowledge of the vertical reference system [7]. One also needs to know that the measurements were taken at a *fixed location* (the gauge's position) and that each height measurement was *controlled by a specific time instant* [7]. This idea was originally introduced by Sinton as a model for different types of thematic cartography [43] and has since been adopted in the context of spatial questions [27, 39]. [41] introduced an *experimental model* that formalizes Sinton's idea from an explicitly experimental perspective. The aim is not only to account for geodata models [39], but more fundamentally, to make explicit the experimental structure underlying any form of geographic information, regardless of how it is represented. In this approach, $exp : X, Y \rightarrow Z$ denotes an experimental function that *fixes* situation variables ranging over X, *controls* situation variables over Y, and *measures* situation variables over Z. Applying this function to examples yields: $ekb \in EKB(f : X, c : Y \rightarrow m : Z)$ a *particular experimental knowledge base* with the same structure, resulting from a finite set of experimental performances – that is, from applying *exp* to some controls in domain Y and recording the resulting outcomes in Z. A geo-dataset can represent such a knowledge base, which makes its inherent experimental structure explicit [41]. Core concepts of spatial information [24], for example, can be interpreted as types of experiments [41]. This includes not only *spatial fields* [23] as functions of spatial location (experiments controlled by vectors in a spatial reference frame) [10], a view that underlies many geodata models [4], but also *coverages* (experiments measuring the spatial area covered by some controlled amount, e.g., of landcover, or of animals) and *lattices* (measuring amounts controlled by spatially tessellated regions, e.g., of administrative territories) (cf. [39, 41]). Previous attempts include applying this idea to geographic quantities (including amounts) in order to define the notion of (spatial) *extensivity* [48], which is fundamental for understanding geographic quantities and their cartographic representation [37]. It can be defined by the additivity of experimental sums in controls and measures [48]. However, while [41] laid the conceptual foundations for the present paper, it remains to construct magnitudes, their quantifications and proportions based on amounts in the context of geographic information. Such constructions are the basis for assessing their *compatibility when deriving new quantities* in a type-safe manner.

3 A theory of quantifiable magnitudes and proportions

Magnitudes are not amounts; rather, they are used to *record and compare the size* of amounts while abstracting from their identity – for example, when quantifying a determinate amount of water that has evaporated from a container. Comparisons between magnitudes are formalized using *proportions*. In standard measurement theory [32], however, magnitudes and proportions are typically collapsed into a single rational number scale, and amounts disappear [45]. By separating amounts, magnitudes, and proportions, we distinguish levels

that must be kept apart for valid geographic measurement: amounts can be aggregated and intersected, magnitudes linearly ordered, and proportions used to scale magnitudes. And unlike classical dimensional analysis [2] or modern quantity calculus [11, 9], which presuppose quantities as elements of a given multi-dimensional structure or of some vector space [51, 34], our theory derives the algebra of linear quantities constructively from amount domains via inductive closure under ratios and products. Using Lorenzen's abstraction principle [31, 29, 28], new domains are generated as equivalence classes of equivalence relations specified on previously constructed domains. Each level of abstraction is specified by constructive norms and definitions in domain-restricted first-order logic. Lorenzen's quantifiers are abstractions over domain-indexed conjunctions and disjunctions: $\bigwedge_{x \in D} \varphi(x)$ and $\bigvee_{x \in D} \varphi(x)$ denote universal and existential quantification over elements x of domain D, while $\wedge$, $\vee$, and $\rightarrow$ denote conjunction, disjunction, and implication. The theory is provably consistent and theorems were proven semi-automatically[8].

3.1 Constructing magnitude scales from amounts

Our reconstruction begins with the establishment of size equivalences over different domains of amounts [19], and then establishes the domains of magnitudes based on such equivalences.

Amount domains

Following methodical constructivist proposals [31, 18], amounts can be reconstructed either *bottom-up*, starting from the smallest atoms and using *set-abstraction* [31], or *top-down*, i.e., by *experimentally controlling* the stuff from which amounts are formed [18].

In top-down construction, amounts are identified by *homogenization norms* and *equivalences* (cf. the research program of *proto-physics* [18]), where different norms result in different kinds of amounts. For *material* amounts (substances), techniques such as sieving, sedimentation, stirring, or measurement are used to separate, mix, or standardize portions. For amounts of *mass*, equivalence among material amounts is established via standard weights and balances, thereby defining mass via observable pull-equivalence using, e.g., rope balances [19]. For *energy* – whether thermal, electrical, or kinetic – amounts are homogenized by bringing container objects into equilibrium by controlling energy transformation (work, heat, motion). For *money or economic value*, amounts are socially constructed through actions, namely (market) exchange trade [31]: goods or services are treated as equivalent if they can be exchanged at agreed ratios, thereby constructing amounts of monetary value. Because these kinds of quantities are constructed from *homogenization norms* and *equivalence relations* rather than from smallest elements, they are considered *atomless*. Any given amount can, practically, be divided further without necessarily reaching the smallest possible unit.

For *bottom-up* constructions of amounts, atoms are explicitly given, because amounts are reconstructed as sets of particular kinds of elements. Each element corresponds, by construction, to a singleton set, forming some smallest part in the partonomy of sets reconstructed from those elements by means of *set abstraction* [30, 31]. In this way, we can reconstruct *spatial regions* as (regular) sets of spatial vectors within a spatial reference frame [20]. We construct *time intervals* as (regular) sets of temporal references (intervals) based on events or processes (e.g., flood intervals) [17]. Finally, we define *amounts of objects* as sets either based on their properties or their relations. For example, amounts of elderly people are defined by age (those whose age is above a threshold), and amounts of citizens are defined by role. Still, the atomic amount of persons always consists of precisely one person.

[8] A Mace4/Prover9 (`https://www.cs.unm.edu/~mccune/prover9/`) script which is called (`quantity_amount_magnitude.txt`) is available under `https://zenodo.org/records/20753038`.

Size equivalences and amount multiples

From a constructivist perspective, sizes are abstractions of *size equivalence relations* in terms of equivalence classes [31, 18]. In top-down constructions of material amounts, such equivalences are established experimentally: *mass equivalence* via rope balances [19], *energy equivalence* via energy transformations, or *value equivalence* via market exchange. As for *bottom-up constructions*, equivalences depend on the kind of set construction involved. For finite sets – such as *amounts of objects* in population counts – size is determined by *cardinality*. However, this strategy fails for (potentially) infinite sets, such as amounts of space or time. In these cases, sizes are determined by comparison with *standard linear, areal, or volumetric forms*, depending on dimensionality. For example, reconstructing the size of a two-dimensional region requires standard geometric forms that *partition* that region. In GIS, e.g., measuring area can be done using Gauss' *shoelace formula*, which presupposes that a region can be decomposed into *triangles* or *trapezoids* [3]. Alternatively, regions may be approximated using grid cells of fixed size. Similarly, *durations* are obtained by partitioning time intervals of equal duration [17]. Essentially, all these methods rely on the ability to determine equivalent sizes, governed by the following norm:

▶ **Constructive Norm 1** (size equivalence). $\bigwedge_{\in A} a, b, c.$

$a \cong_{AE} a$ *(reflexivity)*

$a \cong_{AE} b \to b \cong_{AE} a$ *(symmetry)*

$a \cong_{AE} b \wedge b \cong_{AE} c \to a \cong_{AE} c$ *(transitivity)*

where AE may denote *mass*, *energy*, *value*, *cardinality*, *spatial size* (length, area, volume), or *duration* equivalence. Integers enable the definition of *iterative sums* of amounts. Together with size equivalence, this allows the construction of *multiples*. A multiple [25] is an iterative sum of equally sized[9], mutually non-overlapping amounts:

▶ **Definition 1.** $\bigwedge_{\in A} a, a_1, \ldots, a_n.$

$Multiple^n_{\cong_{AE}}(\{a_1, \ldots, a_n\}, a) \leftrightarrow$

$(\bigwedge_{\in\{a_1,\ldots,a_n\}} x, y.\ x \cong_{AE} y \wedge \neg O(x, y)) \wedge a = ((a_1 + a_2) + \ldots) + a_n$

An amount a is an *n-multiple* if it is the sum of a set of n equal-sized, non-overlapping amounts, where n is the cardinality of the set. For example, a *5-multiple* of the mass of a standard weight is obtained by producing five distinct weights that are mass-equivalent and summing up their amounts. Here, *multiplication* corresponds precisely to the craft of multiplying standard amounts, illustrating that measurement capacities depend on material craftsmanship [18]. In an idealized form, multiples provide the basis for constructing both magnitudes and proportions.

Constructing magnitude domains

Constructing *magnitude domains* requires us to abstract over *size-equivalent amounts*. In this way, the lattice structure of amounts is preserved in part but collapsed into a linear domain corresponding to a *scale*. Consider the *mass magnitude domain*, which is based on an equivalence relation $\cong_{mass}$ between amounts of material determined by a balance [19]. The magnitude domain M_{mass} can then be reconstructed as the set of size-equivalence classes over the amount lattice; each class corresponds to a magnitude in the image of a measurement function μ:

[9] Multiples need to be equally sized to avoid circular reasoning in the construction [20]. Reconstructing quantities from unequally sized amounts by numeric comparison would already presume their quantifiability.

▶ **Definition 2** (magnitude measurement of amounts).
$\mu : A \to M_{AE}$ *with* $\bigwedge_{\in A} x, y.\ x \cong_{AE} y \leftrightarrow \mu(x) = \mu(y)$

Further magnitude properties can be *induced* from amount relations by suitable norms:

▶ **Definition 3** (ordering of magnitudes).
$\bigwedge_{\in A} x, y.\ \mu(x) \leq \mu(y) \leftrightarrow \bigvee_{\in A} z, z'.\ x \subseteq z \wedge z \cong_{AE} z' \wedge z' \subseteq y$

Here, magnitude order is defined by combining amount parthood and equivalence, which requires complete amount lattices (see appendix A). As a constructive norm for magnitude domains, $\cong_{AE}$ must *span* the amount domain so that all amounts become comparable by magnitude. This yields a *total* (linear) ordering:

▶ **Constructive Norm 2** (totality of ordering of magnitudes).
$\bigwedge_{\in A} x, y.\ \mu(x) \leq \mu(y) \vee \mu(y) \leq \mu(x)$

The magnitude domain M_{AE} is equipped with sum and difference operations $+, -$ and neutral elements whose meaning derives from – and preserves – the structure of the amount lattice (a *homeomorphism*):

▶ **Constructive Norm 3** (homeomorphism of magnitudes).
$\bigwedge_{\in A} x, y, z.$

$\mu(0^A) < \mu(1^A)$	*(magnitude bounds)*
$\mu(-x) = (\mu(1^A) + -\mu(x))$	*(magnitude inverse)*
$\mu(x \setminus y) = \mu(x) + -\mu(x * y)$	*(magnitude minus)*
$\mu(x) = \mu(x \setminus y) + \mu(x * y)$	*(magnitude minus neutral)*
$\mu(x + y) = \mu(x) + \mu(y \setminus x)$	*(magnitude addition)*

In mathematical measure theory, such a μ is called a *measure function* [14]. Since we require amount lattices to be non-degenerate (see appendix), it follows that the induced magnitude domain is *non-degenerate*, too:

▶ **Theorem 1** (non-degenerate). $\bigvee_{\in A} x, y.\ (-O(x, y) \wedge \mu(y) = \mu(x))$

That is, there exist at least two non-overlapping amounts of identical magnitude. Furthermore, magnitudes inherit bounds[10], complements, inverses, and an additive identity from their respective amount domains:

▶ **Theorem 2** (Magnitude bounds). $\bigwedge_{\in A} x.\ \mu(0^A) \leq \mu(x) \wedge \mu(x) \leq \mu(1^A)$

▶ **Theorem 3** (Magnitude complement). $\bigwedge_{\in A} x.\ \mu(x) + \mu(-x) = \mu(1^A)$

▶ **Theorem 4** (Additive inverse). $\bigwedge_{\in A} x.\ \mu(x) + -\mu(x) = \mu(0^A)$

▶ **Theorem 5** (Additive identity). $\bigwedge_{\in A} x.\ \mu(x) + \mu(0^A) = \mu(x)$

In this respect, magnitude domains resemble linear vector spaces, though equipped with only a *restricted* form of addition and subtraction:

▶ **Theorem 6** (Restricted addition). $\bigwedge_{\in A} x, y.\ \neg O(x, y) \to \mu(x) + \mu(y) = \mu(x + y)$

▶ **Theorem 7** (Restricted subtraction). $\bigwedge_{\in A} x, y, y'.\ x \subseteq y \wedge y \cong_{AE} y' \to \mu(y') + -\mu(x) = \mu(y \setminus x)$

[10] Note that these bounds are not absolute because they depend on the construction of the underlying amount domain. As long as larger amount domains can be constructed, there are no absolute magnitude limits. Nevertheless, every particular amount domain is bounded by construction and therefore induces bounded magnitudes.

Consequently, no magnitude exceeds $\mu(1^A)$ or falls below $\mu(0^A)$; these bounds are inherited from the amount domain. For example, no mass magnitude can be smaller than the zero amount (0^A). To *quantify magnitudes* on ratio scales and construct *proportions*, however, we need to express magnitudes as numerical multiples of others. This involves *idealizing* [29, 31] multiples using numbers.

3.2 Quantifying magnitudes

In our framework, numbers are not quantities themselves. Rational numbers can be reconstructed inductively from integers[11] in a way that is independent of amounts (cf. [30]). In the following, we will assume fractions $\frac{n'}{n''}$ to represent rational numbers ($\mathbb{Q}$) constructed from integers $n', n'' \in \mathbb{Z}$ in the usual way. Magnitudes are quantified by multiplying and dividing their underlying amounts by *base units*. Integer multiplication is defined as n-fold magnitude addition:[12]

▶ **Definition 4** (integer multiplication of magnitudes). $\bigwedge_{\in M_{AE}} m.\ \bigwedge_{\in I} n$
$m \times n = (((\mu(0^A) + m)_1 + m)_2 \cdots + m)_n$

Thus m is added n times. This extends restricted magnitude addition by allowing for the unrestricted generation of multiples (cf. Sec 3.1). It follows that $m \times 0 = \mu(0^A)$ and $\mu(0^A) \times n = \mu(0^A)$. For positive m and n, integer multiplication is injective:

▶ **Theorem 8** (injectivity of integer multiplication). $\bigwedge_{\in M_{AE}} m, m', m''.\ \bigwedge_{\in I} n, n'$
$(0 < n, n' \wedge \mu(0^A) < m, m', m'') \rightarrow$
$(m' \times n = m \wedge m'' \times n = m \rightarrow m' = m'') \wedge (m' \times n = m \wedge m' \times n' = m \rightarrow n = n')$

Because of injectivity, integer division can be defined as the inverse of multiplication:

▶ **Definition 5** (integer division of magnitudes). $\bigwedge_{\in M_{AE}} m.\ \bigwedge_{\in I} n.$
$(0 < n \wedge \mu(0^A) < m) \rightarrow m \times n^{-1} = \iota_{\in M_{AE}} m'.\ m' \times n = m$

(Hereafter, this *positivity condition* is left implicit for all m occurring in formulas with magnitude quantification.)

Rational multiplication combines division and multiplication:

▶ **Definition 6** (rational multiplication of magnitudes). $\bigwedge_{\in M_{AE}} m.\ \bigwedge_{\in I} n_1, n_2$
$m \times \frac{n_1}{n_2} = (m \times n_2^{-1}) \times n_1$

For example, $1\,kg \times \frac{3}{4}$ is practically obtained by dividing an underlying amount of magnitude 1 kg into four equivalent parts and summing up three equivalent amounts. We require all positive magnitudes to be *quantifiable*: there must be a base magnitude from which these magnitudes can be regenerated by multiplication with a unique rational number.

▶ **Constructive Norm 4** (quantifiability of magnitudes). $\bigvee_{\in M_{AE}} m_{base}.$
$\bigwedge_{\in M_{AE}} m_1, m_2.\ \bigvee_{\in I} n_1, n_2, n_3, n_4.$
$m_{base} \times \frac{n_1}{n_2} = m_1 \wedge m_{base} \times \frac{n_3}{n_4} = m_2 \wedge (\bigwedge_{\in I} n_1, n_2, n_3, n_4 . m_{base} \times \frac{n_1}{n_2} = m_{base} \times \frac{n_3}{n_4} \rightarrow \frac{n_1}{n_2} = \frac{n_3}{n_4})$

[11] And in a similar way, irrationals, differentials, and infinitesimals can be reconstructed from rational numbers, cf. Lorenzen [30].
[12] To avoid confusion, we use "×" instead of "*" to distinguish multiplication from intersection.

This guarantees (a) the existence of a common base m_{base} for all magnitudes of a domain, (b) there is a unique rational quantification of all magnitudes relative to that base, and (c) the transformability of any two magnitudes via the ratio of their quantifications. The *Archimedean property* for ratio scales [32] follows: any non-zero magnitude can be obtained from any other by multiplication with a suitable rational number. This is not an independent axiom but a consequence of quantifiability and the existence of magnitudes greater than $\mu(0^A)$. To quantify a magnitude is to assign the unique rational number required to obtain it from a chosen base:

▶ **Definition 7** (quantification using base units). $\bigwedge_{\in M_{AE}} m, m_{base}.$
$qu_{m_{base}}(m) = \iota_{\in R} \frac{n_1}{n_2}.\ m_{base} \times \frac{n_1}{n_2} = m$

Hence, every magnitude can be decomposed into a multiple of some base unit:

▶ **Theorem 9** (magnitude unit decomposition). $\bigwedge_{\in M_{AE}} m, m_{base}.$
$m = m_{base} \times qu_{m_{base}}(m)$

What we call base units are the same thing as *scale units* (e.g., 1 kg, 1 $meter$). Multiple bases may exist, yielding different but semantically equivalent *measurement scales* (kg, lb, etc.). In bottom-up constructions of finite sets, a natural base is the magnitude of singleton sets. For example, population numbers use the magnitude of one person as base. All bases, like all quantifiable magnitudes, must exceed $\mu(0^A)$.

3.3 Magnitude ratios and products as linear quantities

Pairs of magnitudes under division $(m_1 : m_2)$ are *ratios* and pairs under multiplication $(m_1 \times m_2)$ are *products*. Note that products and ratios *are not magnitudes*, just like population density is not the same kind of quantity as a population number. We generalize magnitudes with their iterated ratios and products under the notion of *linear quantities*[13], forming a domain Q, which is inductively generated from magnitude domains by repeated formation of ratios and products:

▶ **Constructive Norm 5** (ratios, products and linear quantities). $\bigwedge_{\in M_{AE1}} m_1.\ \bigwedge_{\in M_{AE2}} m_2.$
$m_1, m_2 \in Q \wedge \bigwedge_{\in Q} q_1, q_2.\ (q_1 \times q_2) \in Q \wedge (q_1 : q_2) \in Q$

Q must furthermore be constructed in a way such that multiplication is symmetric and associative, admits neutral elements $(m : m)$, and allows the inversion of ratios:

▶ **Constructive Norm 6** (multiplication of linear quantities). $\bigwedge_{\in Q} m_1, m_2, m_3.$

$m_1 \times m_2 = m_2 \times m_1$	*(symmetry)*
$(m_1 \times m_2) \times m_3 = m_1 \times (m_2 \times m_3)$	*(associativity)*
$m_1 \times (m_2 : m_2) = m_1$	*(neutral element)*
$m_1 \times (m_2 : m_3) = (m_1 \times m_2) : m_3$	*(ratio associativity)*
$m_1 : (m_2 : m_3) = m_1 \times (m_3 : m_2)$	*(ratio inversion)*

These *rewriting rules* make $(Q, \times)$ a commutative Abelian group. Standard simplifications of ratios follow from this:

▶ **Theorem 10** (simplifying ratios). $\bigwedge_{\in M_{AE1}} m_1.\ \bigwedge_{\in M_{AE2}} m_2.\ \bigwedge_{\in M_{AE3}} m_3.$
$(m_1 : m_2) : m_3 = m_1 : (m_2 \times m_3)$
$(m_1 \times m_2) : m_1 = m_2$

[13] Ratios and products are called linear quantities because they can be totally ordered inductively, based on the ordering of their constituting magnitudes.

Thus linear quantities can be algebraically reduced whenever their constituent magnitudes *cancel out*. Unlike rational numbers, however, products and ratios of magnitudes need not themselves yield magnitudes; only cancellation reduces them back to a particular magnitude.

Quantification of linear quantities and unit decomposition

Ratios from different magnitude domains (e.g., persons per area) are *heterolithic*; those from a single domain are called *monolithic*. How can we quantity such ratios? In general, rational multiplication of linear quantities is done componentwise:

▶ **Constructive Norm 7** (multiplication of ratios and products). $\bigwedge_{\in Q} m_1.\ \bigwedge_{Q} m_2.\ \bigwedge_{\in I} n_1, n_2.$
$(m_1 : m_2) \times \frac{n_1}{n_2} = ((m_1 \times \frac{n_1}{n_2}) : m_2) = (m_1 : (m_2 \times \frac{n_2}{n_1}))$
$(m_1 \times m_2) \times \frac{n_1}{n_2} = (m_1 \times \frac{n_1}{n_2}) \times m_2$

Combining unit decomposition with ratio and product multiplication then yields:

▶ **Theorem 11** (Linear quantity unit decomposition). $\bigwedge_{\in M_{AE1}} m_1.\ \bigwedge_{\in M_{AE2}} m_2.\ \bigwedge_{\in M_{AE3}} m_3.$
$m_1 \times (m_2 : m_3) = (m^1_{base} \times (m^2_{base} : m^3_{base})) \times (qu_{m^1_{base}}(m_1) \times \frac{qu_{m^2_{base}}(m_2)}{qu_{m^3_{base}}(m_3)})$

Multiplication thus separates into: (i) base-unit algebra (where cancellation of base magnitudes may occur), and (ii) ordinary rational arithmetic on their quantifications. Cancellation rules yield special cases, such as monolithic multiplication:

▶ **Theorem 12** (monolithic multiplication). $\bigwedge_{\in M_{AE1}} m_1.\ \bigwedge_{\in M_{AE2}} m_2, m_3.$
$m_1 \times (m_2 : m_3) = m^1_{base} \times (qu_{m^1_{base}}(m_1) \times \frac{qu_{m^2_{base}}(m_2)}{qu_{m^2_{base}}(m_3)})$

Heterolithic multiplication transforms magnitudes across domains whenever units cancel; otherwise compound units remain:

▶ **Theorem 13** (heterolithic multiplication). $\bigwedge_{\in M_{AE1}} m_1. \bigwedge_{\in M_{AE2}} m_2, m_3.$
$m_2 \times (m_1 : m_3) = m^1_{base} \times (qu_{m^2_{base}}(m_2) \times \frac{qu_{m^1_{base}}(m_1)}{qu_{m^2_{base}}(m_3)})$

For example, $100km^2 \times (60\text{person} : 1km^2) = 1km^2 \times (1\text{person} : 1km^2) \times 100 \times \frac{60}{1} = 6000$ person. We can use heterolithic ratios to quantify important types of proportions, e.g., *density*, in terms of fractions (60 people per square km), and to *transform magnitudes from one domain to another* (here: from areal sizes to number of people).

Equivalence and invariance of ratios

Ratios are equivalent if they only differ by a rational factor in numerator and denominator:

▶ **Definition 8** (equivalence of ratios). $\bigwedge_{\in M_{AE1}} m_1, m_3.\ \bigwedge_{\in M_{AE2}} m_2, m_4.$
$(m_1 : m_2) \cong_{ratio} (m_3 : m_4) \leftrightarrow$
$\bigvee_{\in I} n_1, n_2.\ (m_1 \times \frac{n_1}{n_2}) : (m_2 \times \frac{n_1}{n_2}) = (m_3 : m_4)$

It follows that multiplication with ratios is invariant under equivalent ratios and under changes of base units:

▶ **Theorem 14** (invariance of ratio multiplication). $\bigwedge_{\in M_{AE}} m, m_1, m_2, m'_1, m'_2.$
$(m_1 : m_2) \cong_{ratio} (m'_1 : m'_2) \rightarrow m \times (m_1 : m_2) = m \times (m'_1 : m'_2)$

Proof. All equivalent ratios differ only by a rational factor $\frac{n_1}{n_2}$ in the numerator and the denominator. This means that their quantifications also differ by a constant rational factor, and so this factor cancels out by the rules of rational number fractions. Suppose we use two different constructive base units for quantifying these two ratios. Note that by constr. norm 4, one needs to be expressible as a rational multiple of the other. So the ratios differ only by a constant rational factor in the numerator and denominator, too, and the same argument above applies. ◀

Hence magnitude-ratio multiplication depends only on the proportion represented, not on particular ratio pairs or chosen units.

3.4 Proportions

Theorem 14 allows us to define a domain of *proportions* as equivalence classes of magnitude ratios:

▶ **Definition 9** (proportions of magnitude ratios). $\mu : M_{A'E} \times M_{A''E} \to (M_{A'E} : M_{A''E})$
$\bigwedge_{\in M_{A'E}} x, v. \bigwedge_{\in M_{A''E}} y, w.\ (x : y) \cong_{ratio} (v : w) \leftrightarrow \mu(x : y) = \mu(v : w)$

Proportion domains $M_{A'E} : M_{A''E}$ thus represent ratios independently of any chosen base or equivalent scale. Multiplication of magnitudes with proportions follows the same rules as for ratios, and domains of magnitude products with proportions $M_{A'E} \times (M_{A''E} : M_{A'''E})$ obey the same combination and cancellation principles. This construction parallels the *Eudoxean theory of ratios* [15, book V].

3.5 Types of quantity domains

Following our constructions introduced above, monolithic proportions arise within a single amount domain, e.g., population percentages or shares of energy. Heterolithic proportions span distinct domains, e.g., population density (persons per area), and require separate treatment for the numerator and denominator when computing with them. Based on this,

Table 2 Monolithic and heterolithic proportions. The sign " is the ditto mark. It denotes that the previous type (in the numerator) is repeated in the denominator.

Proportion variant	Type	Num-amount	Denom-amount	Example
Monolithic	Continuous share	Atomless	"	Renewable energy fraction
	Discrete share	Objects	"	Fraction of elderly persons
	Space proportion	Space	"	Fraction of urban land
	Time proportion	Time	"	Commuting time fraction
Heterolithic	Density	Any	Space	Population density
	Speed	Space	Time	Travel speed
	Power	Energy	Time	Energy per hour
	PPer object	Any	Objects	CO_2 per person
	PPer unit	Any	Atomless	Price per kg

we introduce the following terminology (cf. Table 2): *Monolithic proportions* of atomless amounts are called *continuous shares*; object counts yield *discrete shares*. Fractions of space or time yield *space* or *time proportions*. *Heterolithic proportions* generalize familiar spatio-temporal measures: *density*, *flow*, *speed*, *tempo*, and per-object or per-unit rates.

Table 3 Grammar for a terminology of quantity domains.

Types of quantity domains	Abbreviations
object ::= person | panel | neighborhood | ...	*region ::= A(2d)*
atomless ::= material | mass | energy | ...	*interval ::= A(time)*
space ::= 1d | 2d | 3d	*number ::= MA(object)*
atomic ::= space | time | object | ...	*volume ::= MA(3d)*
amount ::= A(atomic | atomless)	*area ::= M(region)*
magnitude ::= M(amount)	*length ::= MA(1d)*
proportion ::= P(magnitude : magnitude)	*duration ::= M(interval)*
cshare ::= P(MA(atomless) : MA(atomless))	*density ::= MA(atomic | atomless) : MA(space)*
dshare ::= P(MA(atomic) : MA(atomic))	*speed ::= length : duration*
	power ::= MA(energy) : duration
	avg ::= MA(atomic | atomless) : number

Table 3 summarizes the types of quantity domains we use. These distinctions capture practically relevant consequences: atomic/object-based amounts have minimal countable units restricting fractions, while atomless amounts do not. Explicit incorporation of units enables us to distinguish *unit conversions* (km → miles) from *quantity transformations* (density → population), supporting transparent, type-safe spatio-temporal modeling.

4 Inference and compatibility of quantifiable experiments

So far, we have treated quantities as independently measurable. In practice, every measurement depends on an experimental setup relating amount domains. By re-embedding quantities into a prior experimental framework [41] (see Sec. 2.2), we can define both the procedures used to generate measurements and how new experiments can be inferred from existing ones. *Quantifiable experiments* are those that fix, control, or measure *amounts*. These amounts can be measured as magnitudes, which can in turn be quantified using rational numbers. The *semantics of GIS operations* for quantities can be understood in terms of *experimental inferences* [41]. Furthermore, whether a set of experimental knowledge bases is *compatible* with respect to a goal experiment reduces to the question whether there is an inference that produces the goal from the sources.

4.1 Experimental combinations

▶ **Definition 10** (experimental combination). $\bigwedge_{\in EKB(f:x,c:Y\to m:M_{A'E})} ekb_1.$
$\bigwedge_{\in EKB(f:x,c:Y\to m:M_{A''})} ekb_2.\bigwedge_{\in EKB(f:x,c:Y\to m:M_{A'''})} ekb_3.$
$ekb_1 \times (ekb_2 : ekb_3) \in EKB(f : x, c : Y \to m : M_{A'E} \times (M_{A''E} : M_{A'''E}))\wedge$
$\bigwedge_{\in Y} y.\bigwedge_{\in M_{A'}} m'.\bigwedge_{\in M_{A''}} m''.\bigwedge_{\in M_{A'''}} m'''.$
$[x, y, (m' \times (m'' : m'''))] \in ekb_1 \times (ekb_2 : ekb_3)) \leftrightarrow ([x, y, m'] \in ekb_1 \wedge [x, y, m''] \in ekb_2) \wedge$
$[x, y, m'''] \in ekb_3)$

This definition captures how new experimental recordings can be inferred by combining existing experimental knowledge bases, matching controls, and combining measurements via magnitude multiplication. For example, consider multiplying a database of region sizes: $rs \in EKB(f : time, c : region \to m : area)$ with a database of regional mass densities $mp \in EKB(f : time, c : region \to m : P(MA(mass) : area))$ at a fixed time t. The latter could represent vegetation mass density in kg/km^2 for each region of a spatial

lattice. If the controlled regions in both databases share a common subset, $rs \times mp$ yields a non-empty database of proportional mass magnitudes for this subset: $rs \times mp \in EKB(f : time, c : region \to m : MA(mass))$. If additionally the measurement units of area are compatible, they cancel out, allowing the mass to be quantified as a rational number: $(0.5km^2 \times (1000kg : 1km^2) = (1kg \times 1000) \times ((1km^2 \times \frac{1}{2}) : 1km^2) = 1kg \times (1000 \times \frac{1}{2}) = 500kg)$. This outcome can be inferred purely from the types of domains and the units of quantification.

Experimental combinations correspond to a generalized form of *local map algebra* [47], allowing computations with different magnitude and proportional units, as well as various experimental control domains beyond raster cells.

4.2 Experimental transformations

In case either the control regions or the measurement units are *incompatible*, one first would need to apply some *transformation* to the experiments in order to render them compatible with each other. The simplest transformation is a *unit conversion*. Since every base magnitude m''_{base} of some domain must be a rational multiple of every other base unit m'_{base} of the same domain, we can always convert units into each other: $m'_{base} \times qu_{m'_{base}}(m''_{base}) = m''_{base}$. The rational number $qu_{m'_{base}}(m''_{base})$ can then be used to transform different quantifications of the same magnitude into each other. For example, suppose the area is given in m^2 instead of km^2, then $qu_{1m}(1km) = \frac{1000}{1}$. More complex transformations may occur in a GIS [44, 7].

5 Case study: estimating the solar potential on roofs in Amsterdam

To evaluate and test the proposed constructive formalism about geographic quantities, we apply it to a case study.[14] Suppose your task as a GIS analyst is to estimate the solar energy potential and its current exploitation on roofs in the city center of Amsterdam, using currently available geodata sources. We have four questions *for each neighborhood* in the city center of Amsterdam, to be answered for the year 2022 (a common fix):

(Q1) *What is the potential amount of solar panels that could be installed on roofs?* $(EKB(f : 2022, c : A(space) \to m : number))$

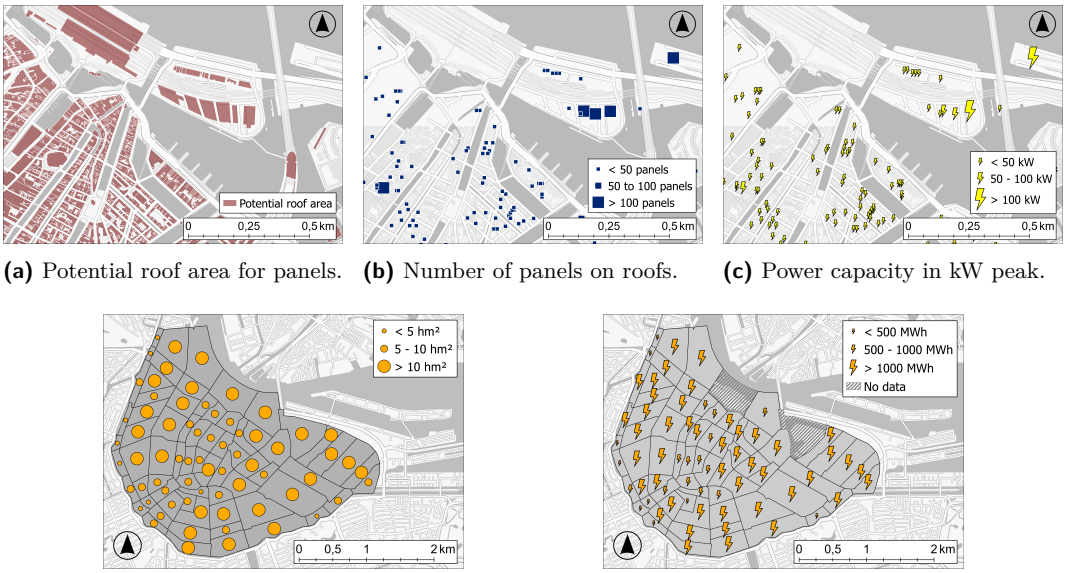
(Q2) *Which proportion of the potential of solar panels is currently installed?* $(EKB(f : 2022, c : A(space) \to m : P(number : number)))$

(Q3) *What is the current amount of solar energy produced?* $(EKB(f : 2022, c : A(space) \to m : M(energy)))$

(Q4) *What proportion of the currently used electricity is produced by solar panels?* $(EKB(f : 2022, c : A(space) \to m : P(M(energy) : M(energy))))$.

Q1 and Q3 ask for magnitudes: the first counts an object amount (number of panels), the second an energy magnitude (solar energy production per year). Q2 and Q4 ask for proportions: the former is a discrete share (installed panels relative to potential panels), the latter a continuous energy share (solar production relative to total electricity consumption). We aim to *infer* which combinations of available sources can answer these questions.

[14] The ArcGIS workspace, a proof script, and related files are available on Zenodo with DOI: 10.5281/zenodo.20753038.



(a) Potential roof area for panels. **(b)** Number of panels on roofs. **(c)** Power capacity in kW peak.

(d) Area of neighborhoods in Amsterdam. **(e)** Electric energy consumption per year.

Figure 1 Various thematic source maps of geographic quantities in Amsterdam, the Netherlands, 2022. Data sources: Amsterdam Data Lab and CBS.

5.1 Geodata sources and their interpretation as quantifiable experiments

To answer these questions, we will reuse data sources from two geodata repositories (illustrated in Fig. 1): the Centraal Bureau voor de Statistiek (CBS)[15] and the data portal of the Municipality of Amsterdam.[16]

(a) *Potential roof area for solar panels*[17]. Assesses the potential roof area suitable for placing solar panels from a spatial and cultural-historical perspective, taking into account the value of building monuments (Fig. 1a). Data format is vector polygons with their size in km^2. This gives an EKB that controls for amounts of space, fixes the year 2022, and measures their area: $roof\text{-}ekb^{area}_{A(space)} \in EKB(c{:}A(space),\ f{:}interval \to m{:}area)$.

(b) *Solar panels*[18]. This dataset represents the spatial distribution of existing solar panels on rooftops in Amsterdam, including their number and power capacities (Fig. 1b and Fig. 1c). The data are vector points with two attributes: panel count and power capacity in kilowatt peak (kW_{peak}).[19]. The EKBs have controls for amounts of panels and measure the spatial area, the number of panels, and power capacity:

$$ekb^{A(space)}_{A(panel)} \in EKB(c{:}A(panel),\ f{:}interval \to m{:}A(space))$$

$$ekb^{number}_{A(panel)} \in EKB(c{:}A(panel),\ f{:}interval \to m{:}number)$$

$$ekb^{power}_{A(panel)} \in EKB(c{:}A(panel),\ f{:}interval \to m{:}power)$$

[15] The neighborhood statistics ("Kerncijfers wijken en buurten 2022"): `https://dataportal.cbs.nl/detail/CBS/85318NED`. Date accessed: 11th Feb 2026.
[16] `https://maps.amsterdam.nl`
[17] `https://maps.amsterdam.nl/zonnedaken_erfgoed/`. Date accessed: 11th Feb 2026.
[18] `https://maps.amsterdam.nl/zonnepanelen/`. Date accessed: 11th Feb 2026.
[19] Kilowatt peak denotes the power output of solar panels under standardized ideal insolation conditions.

From an experimental perspective, the first constitutes an example of a *coverage*, in which space is used as a measure rather than as a control.

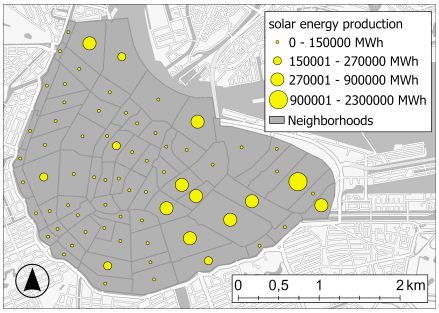
(c) *Neighborhoods*[20]. Represents the regions of administrative units on the "Buurt" level with quantities that are statistical aggregations over these regions (Fig. 1d, Fig. 1e), including their areas in hm^2 and the electric energy consumption of households in MWh.

$$\textit{neigh-ekb}^{M(energy)}_{A(space)} \in EKB(c{:}A(space),\ f{:}interval \to m{:}M(energy))$$

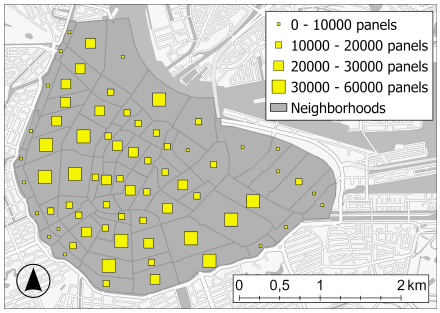
$$\textit{neigh-ekb}^{area}_{A(space)} \in EKB(c{:}A(space),\ f{:}interval \to m{:}area)$$

From an experimental perspective, these are examples of a spatial lattice, where we control tessellated amounts of space via administrative objects and measure various magnitude sums over space.

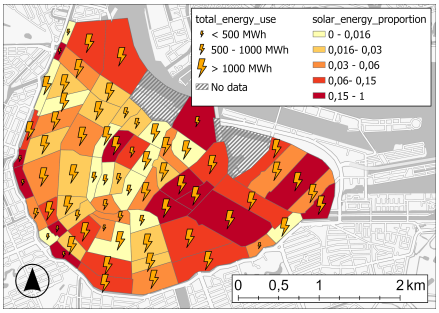
These sources differ not only in the types of quantities but also in their spatial support units, which serve as experimental controls. We shall therefore reason about the compatibility of both the quantities and their units, as well as their experimental structures (see Fig. 3).



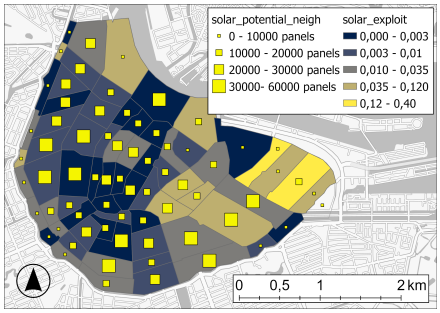
(a) Solar energy production per year in MWh.



(b) Solar energy potential in number of panels.



(c) Solar energy proportion (fraction of energy use).



(d) Solar panel exploitation (fraction of potential).

Figure 2 Thematic maps of derived quantities assessing the solar energy production, the solar potential, and its exploitation on roofs in Amsterdam neighborhoods.

5.2 Compatibility and inference of quantity types for example maps

We showed how quantities can be transformed and inferred to answer the four questions. A key challenge is that the source and target quantities use different controls: potential roof areas and installed rooftop panels do not share controls with the neighborhood-based target

[20] `https://www.cbs.nl/nl-nl/visualisaties/cijfers-op-de-kaart`. Date accessed: 11th Feb 2026.

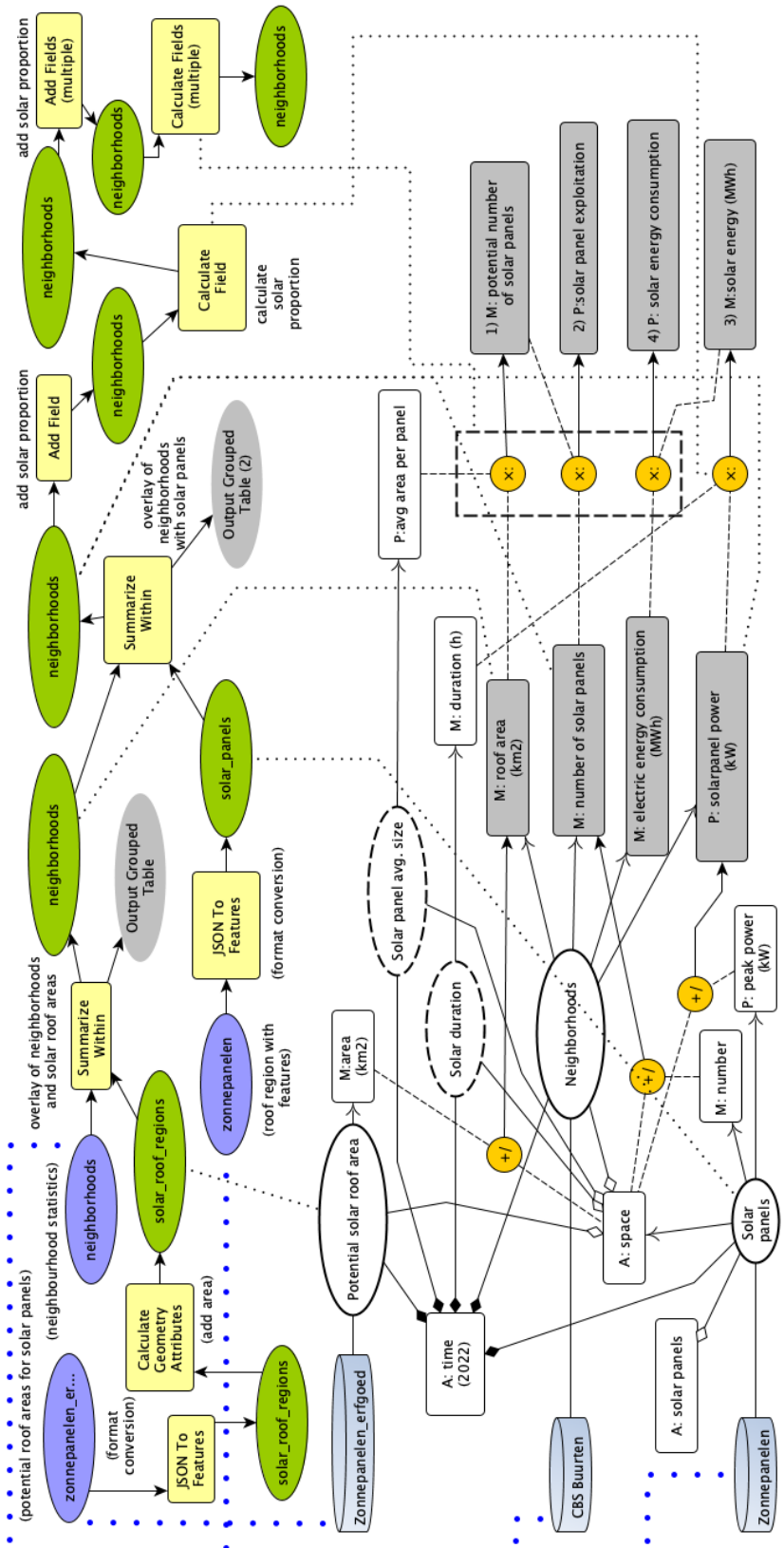


Figure 3 Experimental inference of geographic quantities for the solar energy scenario and its correspondence with a GIS workflow.
The upper part is the workflow of data processing in ArcGIS. The lower part is the modelling of quantities. Blue dotted lines are about the resource-data correspondence. The black dotted lines between the two parts are about the variable and operations correspondence. For the upper part, the ellipses nodes denote data and the boxes denote data processing (data format conversion, summarization, calculation, etc). Grey ellipses are for outputs. For the lower part, thick ellipses denote EKBs. Dashed ellipses denote (implicit) attributes. Arrows with black diamonds indicate fixes, those with white diamonds represent controls, and curly arrows correspond to measures. White boxes are primary source quantities, and grey boxes denote derived quantities. The abbreviations $A:$, $M:$, $P:$ stand for *amount of*, *magnitude* and *proportion*. Orange circles denote experimental transformations $(+/)$ (using sums and differences of either amounts or magnitudes) and combinations $(\times:)$ (using magnitude products and ratios).

experiments. As a result, straightforward experimental combinations or concatenations of these EKBs yield empty results because there are no common controls. Yet, we can use *experimental aggregations* that are based on *amount/magnitude sums and differences* (the orange circles with +/ in Fig. 3) to transform both the solar panels and the potential solar roof area experiments into corresponding neighborhood experiments. Such transformations are implemented as spatial overlay operations in a GIS, which can be used to cut and merge spatial regions into larger ones to fit into "cookie cutter" regions, and to simultaneously sum or subtract magnitude values of their measures. An example would be a *vector overlay* with a *sum function*, such as ArcGIS's *Summarize Within* tool. In the example above, we used this tool to summarize both the potential roof areas and the solar panel point data into neighborhood polygons, yielding three new neighborhood attributes representing: magnitudes of roof area per neighborhood ($\textit{roof-neigh-ekb}^{\textit{area}}_{A(\textit{space})}$), aggregated from $\textit{roof-ekb}^{\textit{area}}_{A(\textit{space})}$; number of solar panels per neighborhood ($\textit{panel-neigh-ekb}^{\textit{number}}_{A(\textit{space})}$), aggregated from $\textit{ekb}^{\textit{number}}_{A(\textit{panel})}$; solar power per neighborhood ($\textit{panel-neigh-ekb}^{\textit{power}}_{A(\textit{space})}$), aggregated from $\textit{ekb}^{\textit{power}}_{A(\textit{panel})}$. Conceptually, these aggregation steps replace object-amount based controls (e.g., panels or roofs) by spatial controls (neighborhood regions) and infer magnitude sums.

Recall that our EKBs are conceptual models for interpreting geodata, not data models (in Fig. 3, we distinguish between EKBs and actual data sources), since not all aspects needed for interpretation are explicitly represented in the data. Power is defined as a proportion of magnitude of energy per duration. To convert power back into solar energy, we need to have information about the duration of insolation within the respective control region for the fixed year, which can be obtained from weather data. We assume 4 hours of sunshine per day, i.e. $365 \times 4 = 1460$ sun hours per year. Thus, an entire *separate experiment* (*solar duration*, see Fig. 3) hides behind these simple numbers, which we represent by: $\textit{solar-neigh-ekb}^{\textit{duration}}_{A(\textit{space})}$. Similarly, to estimate a potential number of solar panels, we need to combine the aggregated roof area with a panel size, which could be obtained from another data source. Yet, this case study adopts a rough estimate of $2m^2$ per panel. We represent this implicit knowledge by: $\textit{panel-neigh-ekb}^{P(\textit{area:number})}_{A(\textit{space})}$. The remaining steps to derive the target maps are based on *experimental combinations* of magnitudes and proportion measures using the following magnitude multiplications (the orange circles with $\times$: in Fig. 3), including quantifications and necessary unit conversions:

1. potential number of solar panels on roofs per neighborhood:
   $\textit{roof-neigh-ekb}^{\textit{area}}_{A(\textit{space})} : \textit{panel-neigh-ekb}^{P(\textit{area:number})}_{A(\textit{space})} = \textit{roof-panel-neigh-ekb}^{\textit{number}}_{A(\textit{space})}$, since
   $\textit{area} : (\textit{area} : \textit{number}) = \textit{area} \times (\textit{number} : \textit{area}) = \textit{number} \times (\textit{area} : \textit{area}) = \textit{number}$
   The roof areas are measured in km^2, while panel sizes are given in m^2. Since $1\,km^2 = 1{,}000{,}000, m^2$, the units are first converted to the same scale and then cancel out, yielding the number of panels.
2. solar exploitation (discrete share of number of panels) per neighborhood:
   $\textit{panel-neigh-ekb}^{\textit{number}}_{A(\textit{space})} : \textit{roof-panel-neigh-ekb}^{\textit{number}}_{A(\textit{space})} = \textit{panel-neigh-ekb}^{P(\textit{number:number})}_{A(\textit{space})}$
3. solar panel energy production per neighborhood:
   $\textit{panel-neigh-ekb}^{\textit{power}}_{A(\textit{space})} \times \textit{solar-neigh-ekb}^{\textit{duration}}_{A(\textit{space})} = \textit{solar-neigh-ekb}^{M(\textit{energy})}_{A(\textit{space})}$, since
   $\textit{power} \times \textit{duration} = (M(\textit{energy}) : \textit{duration}) \times \textit{duration} = M(\textit{energy}) \times (\textit{duration} : \textit{duration}) = M(\textit{energy})$. The resulting energy is obtained in kWh and subsequently converted to MWh by dividing by 1000.
4. solar energy proportion per neighborhood:
   $\textit{solar-neigh-ekb}^{M(\textit{energy})}_{A(\textit{space})} : \textit{neigh-ekb}^{M(\textit{energy})}_{A(\textit{space})} = \textit{solar-neigh-ekb}^{P(M(\textit{energy}):M(\textit{energy}))}_{A(\textit{space})}$

They provide valid answers to our four questions under the practical restrictions of using approximate panel sizes and solar hours.

Our theory also allows for identifying incompatible combinations. For example, multiplying solar power by the number of solar panels or by the roof area instead of solar duration does not yield any meaningful magnitude or proportion. Formally, $power \times area = (M(energy) : duration) \times area = M(energy) \times (area : duration)$ produces a fictive compound unit (e.g., $MWh \times km^2/h$) that not only lacks experimental interpretation, but also cannot be reduced to any magnitude or proportion of interest for answering any of our 4 questions.

6 Conclusion

This paper proposes a constructive semantic theory for determining when geographic quantities can be validly combined. Note that the purpose of our theory does not lie in enabling new kinds of geo-computations. Rather it makes an implicit numeric practice explicit and thereby delivers, for the first time, validity criteria for forming geographic quantities. In contrast to standard measurement theory, we distinguish amounts, magnitudes, and proportions, provide a constructive formalization of these domains, and introduce a notion of compatibility that accounts for the operations available on each domain and the principled derivation of linear quantities from amount domains. The relevance of our theory thereby goes beyond geography, since it enables explicit compatibility constraints for quantities and programmatic validation of numeric workflows. Yet, embedding these quantity types into our experimental framework offers a conceptual lens on GIS operations and their capacity to derive new quantities. The Amsterdam solar case study demonstrated how the theory supports type-safe transformations, unit conversions, and proportional reasoning, thereby establishing a formal basis for assessing quantitative compatibility and the validity of GIS workflows. In increasingly automated AI-driven contexts, our theory might form the backbone of a *neuro-symbolic* AI [16, 33] for valid computing with geographic quantities. Future work should strengthen empirical evaluation, for example by applying the theory to GeoAI or GeoQA benchmark corpora, further investigate its logical properties, extend it to more complex experimental transformations (including concatenations, spatial interpolations, and conversions) in the context of GIS workflow semantics [44], and translate it into metadata standards and algorithms for automated assessment of geodata compatibility in indirect question answering [39].

References

1 B. Aameri et al. The FOUnt ontologies for quantities, units, and the physical world. *Applied Ontology*, 15(3):313–359, 2020. doi:10.3233/AO-200231.

2 P. W. Bridgman. *Dimensional Analysis*. Yale University Press, New Haven, 1922.

3 P. A. Burrough et al. *Principles of geographical information systems*. Oxford Uni Press, 2015.

4 G. Câmara, M. J. Egenhofer, K. R. Ferreira, P. R. de Andrade, G. R. de Queiroz, A. H. Sánchez, J. Jones, and L. Vinhas. Fields as a Generic Data Type for Big Spatial Data. In M. Duckham, E.J. Pebesma, K. Stewart, and A. U. Frank, editors, *Geographic Information Science - 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*, volume 8728 of *Lecture Notes in Computer Science*, pages 159–172. Springer, Springer, 2014. doi:10.1007/978-3-319-11593-1_11.

5 R. Casati et al. *Parts and places: The structures of spatial representation*. MIT Press, 1999. doi:10.7551/mitpress/5253.001.0001.

6 N. R. Chrisman. Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems*, 25(4):231–242, 1998.

7 N. R. Chrisman. *Exploring Geographic Information Systems, 2nd Edition*. Wiley, 2002.

8 E. R. Cohen et al. *Quantities, Units and Symbols in Physical Chemistry, IUPAC Green Book*. Royal Society of Chemistry, January 2007.

9 P. De Bièvre. The beginning of the end of the confusion in concepts and terms? The new international vocabulary of basic and general concepts in metrology (VIM). *Accreditation and quality assurance*, 12(6):279–281, 2007. doi:10.1007/s00769-007-0276-3.

10 A. Galton. Fields and objects in space, time, and space-time. *Spatial cognition and computation*, 4(1):39–68, 2004. doi:10.1207/s15427633scc0401_4.

11 P. Glavič. Quantities and units in chemical and environmental engineering. *Standards*, 2(1):43–51, 2022. doi:10.3390/standards2010004.

12 M. Gruninger, B. Aameri, C. Chui, T. Hahmann, and Y. Ru. Foundational Ontologies for Units of Measure. In S. Borgo, P. Hitzler, and O. Kutz, editors, *Formal Ontology in Information Systems - Proceedings of the 10th International Conference, FOIS 2018, Cape Town, South Africa, 19-21 September 2018*, volume 306 of *Frontiers in Artificial Intelligence and Applications*, pages 211–224. IOS Press, 2018. doi:10.3233/978-1-61499-910-2-211.

13 G. Guizzardi. On the Representation of Quantities and their Parts in Conceptual Modeling. In A. Galton and R. Mizoguchi, editors, *Formal Ontology in Information Systems, Proceedings of the Sixth International Conference, FOIS 2010, Toronto, Canada, May 11-14, 2010*, volume 209 of *Frontiers in Artificial Intelligence and Applications*, pages 103–116. IOS Press, 2010. doi:10.3233/978-1-60750-535-8-103.

14 P. R. Halmos. *Measure Theory*. Springer, 1974. doi:10.1007/978-1-4684-9440-2.

15 T. L. Heath. *The Thirteen Books of Euclid's Elements*. Dover Publications, Inc, 1956.

16 P. Hitzler, M. Ebrahimi, M. K. Sarker, and D. Stepanova. Neuro-symbolic AI and the semantic web, 2024. doi:10.3233/SW-243711.

17 P. Janich. Chronometrie. In *Das Maß der Dinge: Protophysik von Raum, Zeit und Materie*, pages 139–189. Suhrkamp, 1997.

18 P. Janich. *Das Maß der Dinge: Protophysik von Raum, Zeit und Materie*. Suhrkamp, 1997. https://www.suhrkamp.de/buch/peter-janich-das-mass-der-dinge-t-9783518289341.

19 P. Janich. Die Eindeutigkeit der Massenmessung und die Definition der Traegheit: Hylometrie. In *Das Maß der Dinge: Protophysik von Raum, Zeit und Materie*, pages 271–288. Suhrkamp, 1997.

20 P. Janich. Die protophysikalische Begruendung der Geometrie. In *Das Maß der Dinge: Protophysik von Raum, Zeit und Materie*, pages 35–72. Suhrkamp, 1997.

21 K. Janowicz et al. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020. doi:10.1080/13658816.2019.1684500.

22 K. Janowicz et al. GeoFM: how will geo-foundation models reshape spatial data science and GeoAI? *Int. J. of Geographical Information Science*, 39(9):1849–1865, 2025. doi:10.1080/13658816.2025.2543038.

23 K. K. Kemp. Fields as a framework for integrating gis and environmental process models. part 1: Representing spatial continuity. *Transactions in GIS*, 1(3):219–234, 1996. doi:10.1111/j.1467-9671.1996.tb00046.x.

24 W. Kuhn. Core concepts of spatial information for transdisciplinary research. *Int. J. of Geographical Information Science*, 26(12):2267–2276, 2012. doi:10.1080/13658816.2012.722637.

25 M. Lachmair, S. Ruiz Fernández, K. Moeller, H.-C. Nuerk, and B. Kaup. Magnitude or multitude – what counts? *Frontiers in psychology*, 9:522, 2018.

26 R. Lange. Vom koennen zum erkennen. die rolle des experimentierens in den wissenschaften. In *Methodischer Kulturalismus: Zwischen Naturalismus und Postmoderne*, pages 157–196. Suhrkamp, 1996.

27 G. Langran. Producing answers to spatial questions. In *1991 ACSM-ASPRS Annual Convention*, volume 6, pages 133–133, 1991.

28 P. Lorenzen. *Einführung in die operative Logik und Mathematik*, volume 78. Springer, 1955. doi:10.1007/978-3-642-86518-3.

29 P. Lorenzen. *Normative logic and ethics*. Bibliographisches Institut, Mannheim, 1969.

30 P. Lorenzen. *Differential and integral: a constructive introduction to classical analysis*. University of Texas Press, 1971.

31 P. Lorenzen. *Lehrbuch der konstruktiven Wissenschaftstheorie*. Springer, 1987. doi:10.1007/978-3-476-02758-0.

32 R. D. Luce and Suppes P. Representational measurement theory. In *Stevens' Handbook of Experimental Psychology*. John Wiley & Sons Inc, 2002. doi:10.1002/0471214426.pas0401.

33 G. Mai et al. Symbolic and subsymbolic GeoAI: Geospatial knowledge graphs and spatially explicit machine learning. *Trans. GIS*, 26(8):3118–3124, 2022. doi:10.1111/tgis.13012.

34 L. Mari and L Giordani. Quantity and quantity value. *Metrologia*, 49(6):756–764, 2012. doi:10.1088/0026-1394/49/6/756.

35 B. Plewe. A Case for geographic masses. In S. Timpf, C. Schlieder, M. Kattenbeck, B. Ludwig, and K. Stewart, editors, *14th International Conference on Spatial Information Theory, COSIT 2019, Regensburg, Germany, September 9-13, 2019*, volume 142 of *LIPIcs*, pages 14:1–14:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.COSIT.2019.14.

36 H. Rijgersberg et al. Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13, 2013. doi:10.3233/SW-2012-0069.

37 S. Scheider et al. Distinguishing extensive and intensive properties for meaningful geo-computation and mapping. *Int. J. of Geographical Information Science*, 33(1):28–54, 2019. doi:10.1080/13658816.2018.1514120.

38 S. Scheider et al. Pragmatic GeoAI: Geographic information as externalized practice. *KI-Künstliche Intelligenz*, 37(1):17–31, 2023. doi:10.1007/s13218-022-00794-2.

39 S. Scheider, R Meerlo, V. Kasalica, and A.-L. Lamprecht. Ontology of core concept data types for answering geo-analytical questions. *J. Spatial Inf. Sci.*, 20(1):167–201, 2020. doi:10.5311/JOSIS.2019.20.555.

40 S. Scheider and J.A. Verstegen. What Is a Spatio-Temporal Model Good For?: Validity as a Function of Purpose and the Questions Answered by a Model. In Adams B., A. L. Griffin, S. Scheider, and G. McKenzie, editors, *16th International Conference on Spatial Information Theory, COSIT 2024, Québec City, Canada, September 17-20, 2024*, volume 315 of *LIPIcs*, pages 7:1–7:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.COSIT.2024.7.

41 S. Scheider and J.A. Verstegen. The Inherent Structure of Experiments as a Constraint to Spatial Analysis and Modeling. In K. Sila-Nowicka, A. Moore, D. O'Sullivan, B. Adams, and M. Gahegan, editors, *13th International Conference on Geographic Information Science, GIScience 2025, Christchurch, New Zealand, August 26-29, 2025*, volume 346 of *LIPIcs*, pages 17:1–17:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.GIScience.2025.17.

42 P. Simons. Density, angle, and other dimensional nonsense: How not to standardize quantity. In *Johanssonian investigations*, pages 516–534. De Gruyter, 2013. doi:10.1515/9783110322507.516.

43 D. Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 1978. URL: https://cir.nii.ac.jp/crid/1571417125890346624.

44 N. Steenbergen et al. Procedural metadata for geographic information using an algebra of core concept transformations. *Journal of Spatial Information Science*, 27:51–92, 2023. doi:10.5311/JOSIS.2023.27.268.

45 P. Suppes and J.L. Zinnes. *Basic Measurement Theory*, volume 1, pages 1–76. John Wiley & Sons, 1963.

46 L. Talmy. *Toward a Cognitive Semantics, Vol I: Concept Structuring Systems*. MIT Press, 2000. doi:10.7551/mitpress/6847.001.0001.

47 D. Tomlin. *Geographic Information Systems and Cartographic Modeling*. Englewoods Cliffs, Prentice-Hall, 1990.

48 E.J. Top et al. The semantics of extensive quantities within geographic information. *Applied Ontology*, 17(3):337–364, 2022. doi:10.3233/AO-220268.

49 E.J. Top et al. Evidence of the use of extensive and intensive quantity concepts in experienced readers of cartographic maps. *Int. J. of Cartography*, pages 1–24, 2025. doi:10.1080/23729333.2025.2471689.

50 L. Travé-Massuyès and N. Piera. The Orders of Magnitude Models as Qualitative Algebras. In N. S. Sridharan, editor, *Proceedings of the 11th International Joint Conference on Artificial Intelligence. Detroit, MI, USA, August 1989*, volume 89, pages 1261–1266. Morgan Kaufmann, 1989. URL: http://ijcai.org/Proceedings/89-2/Papers/066.pdf.

51 H. Whitney. The mathematics of physical quantities: Part II: Quantity structures and dimensional analysis. *The American Mathematical Monthly*, 75(3):227–256, 1968. doi:10.1080/00029890.1968.11970972.

A Detailed preliminaries for amounts

Amount domains constitute non-linearly ordered *additive part–whole structures* closed under sums, subtractions, intersections and unions [48]:

Norms for portioning amounts. We assume a partially ordered algebra, where the symbol $\subseteq$ ranges over amounts of a particular amount domain A (such as sand, gravel, or clay): For all $a, b, c \in A$, we have $a \subseteq a$ (Amount order reflexivity), $(a \subseteq b) \leftrightarrow (b \supseteq a)$ (Amount order gte/lte), $(a \subseteq b \wedge a \supseteq b) \rightarrow a = b$ (Amount order antisymmetry), and $(a \subseteq b \wedge b \subseteq c) \rightarrow a \subseteq c$ (Amount order transitivity). Based on these norms, we define strict order: $(a \subset b) \leftrightarrow (a \subseteq b \wedge \neg(b \subseteq a))$. Strictly ordered amounts are not identical – one is a proper part of the other.

Summing, subtracting, and intersecting amounts. We define "+" for adding two amounts to form a bigger one and "*" for giving only the overlapping (common) part of two amounts (i.e., intersection/product). We have the constructive norm: $\bigwedge_{\in A} a, b, c.\ a+b = b+a$ (Amount sum symmetry), $a*b = b*a$ (Amount prod symmetry), $(a+b)+c = a+(b+c)$ (Amount sum associativity), $(a*b)*c = a*(b*c)$ (Amount prod associativity), $a*(b+c) = (a*b)+(a*c)$, (Amount prod distributivity), $a+(b*c) = (a+b)*(a+c)$ (Amount sum distributivity). The order of summing or intersecting amounts is irrelevant, leading to symmetry and associativity norms as well as distributivity norms. Amounts can also be *subtracted* from each other (symbol $\backslash$), and thereby have certain well defined neutral elements. As introduced in Section 2, the empty amount 0^A ("nothing") is neutral for sums and absorbing for intersections, and the total amount 1^A ("everything") is neutral for intersections and absorbing for sums. If one removes an amount equal to that of itself, nothing remains (i.e., the inverse amount shows $a \backslash a = 0^A$.). Subtracting/adding this empty amount to another does not change it: $b \backslash 0^A = b$ and $a + 0^A = a$. Furthermore, 1^A ($1^A \in A$) is the neutral element for amount products, and it corresponds to "everything" ($1^A \in A$). For amounts $a, b \in A$, removing everything from something leaves nothing ($b \backslash 1^A = 0^A$), and adding it results in everything ($a + 1^A = 1^A$). The product of nothing with anything is nothing ($a * 0^A = 0^A$), while the product of everything with something does not make a difference to something ($a * 1^A = a$). Based on this, we can define overlaps of amounts in terms of non-empty intersections: $\bigwedge_{\in A} a, b, c. O(a, b) \leftrightarrow (0^A \subset a * b)$.

Amount lattices. Amounts also obey clear rules about how parts combine under sums and products, making them, in mathematical terms, *Boolean lattices*.[21] This means, that for $a, b, c \in A$, the existence of joins and meet always exists: $a \subseteq a + b$, and $a * b \subseteq a$, and we

[21] https://www.umsl.edu/~siegelj/SetTheoryandTopology/The_algebra_of_sets.html

always have $(a \subseteq c \wedge b \subseteq c) \rightarrow a + b \subseteq c$ and $(c \subseteq a \wedge c \subseteq b) \rightarrow c \subseteq a * b$. We can always form the *complement* (symbol $-$) of a given amount, which is everything that is not part of that amount. Thus, we have $a + (-a) = 1^A$ and $a * (-a) = 0^A$. The complement of a complement is the amount itself: $-(-a) = a$ (Amount involution). Finally, we have two laws of De Morgan: $-(a + b) = -a * -b$ and $-(a * b) = -a + -b$. It follows by construction that the empty amount needs to be part of every other amount, which means that there are no "negative" amounts: $(0^A \subseteq a)$ for every $a \in A$. This was used to show that magnitudes are also non-negative. We have subtraction defined as $a \backslash b := a * (-(b * a))$ for all $a, b \in A$. An important norm that prevents degenerate amount lattices is that their parts are always *non-linearly ordered*. This means there are always some parts that are not part of each other. Thus, it has the non-totality axiom: $\bigvee_{a,b \in A} \neg(a \subseteq b)$.[22] More details can be found in [48].

[22] In consequence, amount lattices consist of at least two different amounts on the same hierarchical level.

Qualitative Spatial Representation for Material Documents

Mark Bell
The National Archives, Kew, UK

John G. Stell
University of Leeds, UK

Abstract

This paper proposes the use of Qualitative Spatial Representation (QSR) to model the complex relationships of paper documents and books, both in their physical and digital forms. We begin with motivating examples from The National Archives' holdings, and literature, to describe the myriad ways something as simple as text on paper can be manifested, and introduce appropriate terminology. The primary focus is on the challenges of working with documents comprised of multiple pages, and understanding the relationships between the pages themselves, text items which span more than one page. We end by presenting case studies using Allen Intervals to represent relationships between physical documents, their digitised representations, and the text on the page.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning; Applied computing → Document management and text processing; Applied computing → Arts and humanities

Keywords and phrases Qualitative spatial representation, QSR, spatial text, document understanding, information extraction, Allen interval relations, archives

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.3

Funding *John G. Stell*: Research supported by the UK's Economic and Social Research Council (ES/W003473/1) and the US National Science Foundation (NSF#2225179).

1 Introduction

The aim of this paper is to explore the role of Qualitative Spatial Representation (QSR) in the context of information within material documents, such as books and other media based on flat surfaces that can support text whether handwritten or printed.

A conventional book consists of a sequence of pages which follow each other in a straightforward linear order. This sequence in a physical book need not be followed by the reader, who can jump randomly from one page to another, or methodically move backwards and forwards in studying some aspect of the contents. However, the structure of the book itself still appears fundamentally as a sequence, despite our ability to explore in a non-linear fashion. The idea of a text as a space to explore has a form of parallel in the work of Mocnik and Fairbairn [24] who consider the ways maps can be enabled to support story telling. We argue that the interaction between spatial placement on a page and the spatial location of the page among other pages leads to a relatively neglected aspect of qualitative spatial information which is not as limited as it might initially appear, and which has practical applications. We present applications to information extraction from archives of material documents which also exist as digitised images, and applications to other areas too.

Today, books that exist as material copies are frequently experienced digitally as images or other forms of data on a screen. Texts that are "born digital" may only become physical copies as a derived form when printed out and do not initially exist as material objects. In this paper our concern is with material documents, motivated mainly by archival collections but also by contemporary uses of books in which the material is essential, such as "artists'

© Crown copyright. Re-used under the terms of the Open Government Licence v3.0;
licensed under Creative Commons License CC-BY 4.0

books" and some experimental literature. The digital is, however, a key feature of our investigation. In the archive which provides specific examples we discuss below, there is a very substantial body of digitised versions of material documents. As we explain in detail in the paper, the correspondence between the material originals and the digitised versions is spatially complex. This happens because the originals often depend for their interpretation on qualitative spatial relationships involving the material arrangement. For example, that some text or mark is located on the other side of the surface from another. This "other side of" in the context of a leaf in a book or other writing surface is a feature of the space of material documents that is the topic of this paper. It presents a practical challenge and is spatially complex because a digital document, such as a pdf of images of pages, may not explicitly record such spatial structure but it may be possible to deduce the structure from images.

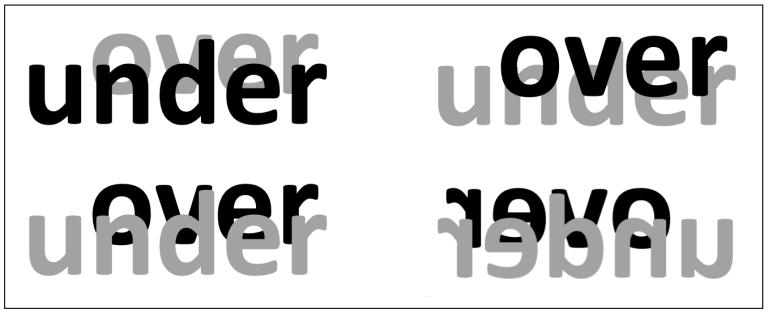


Figure 1 Spatial Relations between apparent layers. In the lower right is over over or under under?

We are mainly focussing our attention on books and documents as sources of information for the reader or viewer. This is significant because some spatial aspects depend critically on there being a point of view, or an observer experiencing the document in a particular orientation. This would not be the case if we only wanted to describe the structure of a book as a three-dimensional construction independent of any special viewpoint. The importance of the viewer can be explained in connection with Figure 1. In the figure are four pairs of words, one grey and one black. Looking at the top row first, the black word appears "on top of" the grey word in each case. We can imagine the words printed on transparent sheets with the sheet carrying the black word uppermost – that is closest to the imagined observer. In the bottom row on the left the grey word is on top. If we had transparent sheets arranged to show the example at the top right, we could obtain the lower left example by swapping the top and bottom sheets. Now consider the lower right example. Is the black word on top of the grey one, or is the grey one on top? There are only two possible answers and either can be argued for.

If we think in terms of transparent sheets bearing the words, then the lower right example would be obtained by turning over, as a single thing, the two sheets from the upper right example. The black word is now physically underneath the grey word. The alternative answer could be justified by saying that the marks constituting the black word need to viewed "from the front", otherwise it is not really a word, and similarly for the grey word. Thus even though to one observer the grey marks might be uppermost, there is an implied position which is actually "from the other side" and from there the black word would be closest to the observer.

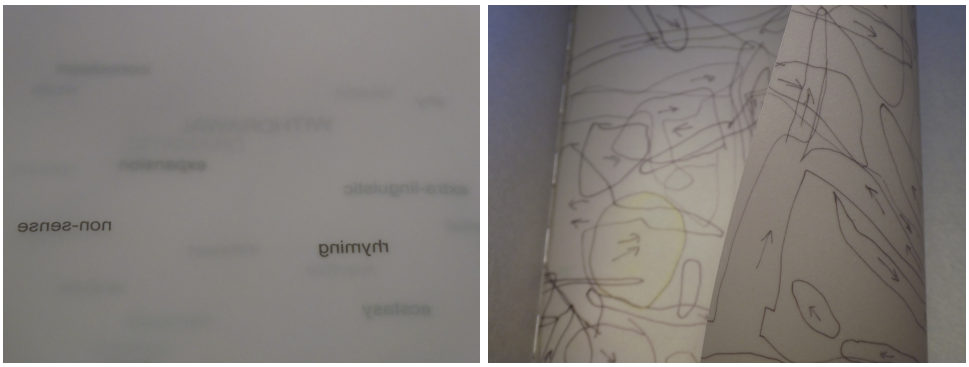


Figure 2 material layers and almost transparency. On the right from [10, p73] with a detail from Mehretu's *Migration Direction Map*, 1996, ink on Mylar. On the left a succession of pages printed on paper that allows a degree of transparency from [4].

Our position in this paper is that either of the answers about which word is "on top" and which is "underneath" is possible and will depend on the particular situation being modelled. Maybe both are needed simultaneously, in which case a formal model needs to incorporate at least one observer as a component in the model. It should be noted that the example here is not unrealistic as our potential applications include examples such as those illustrated in Figure 2 where pages are not opaque. Galleries, archives, libraries and museums have collections which may include artists' books, experimental literature, or, as in the right of Figure 2, catalogues showing drawings made on transparent media that are reproduced on sheets that can be viewed from both sides. Such examples might need to be digitised or might exist within corpora which need to be searched. Search criteria might well need to allow a user to find items where something is "on top of" something else.

To give another example to show how relationships across different surfaces are less than straightforward, consider whether two pages "have the same layout", whether page numbers are "in the same place" on different pages, and whether two pages "coincide exactly". The notion of "the same as" leads to well-known philosophical conundrums involving identity, such as "the ship of Theseus". In our context the issues relate more to geometrical concepts, such as two triangles being "the same" if one can be superposed on the other, which is itself by no means straightforward. Consider a book which has an opening showing two pages in which the page numbers are both positioned in the upper right corner of the page. That is the lefthand page has the page number close to the page edge that is attached in the binding, and the righthand page has the page number close to the edge that is free to move and parallel to the spine of the book. This is different from the conventional traditional layout in which layout on the left is the mirror image of layout on the right, but it does occur in many examples from the 20th century onwards. In one sense the page numbers appear "in the same place" on the two pages. In geometrical terms, by translating the lefthand page to superpose it on the righthand one the positions of the page numbers would coincide. However, if the book were closed, so the left and right pages were in contact, then the page numbers would not coincide. This again, is not an artificial example because experimental literature such as House of Leaves [11] discussed in Section 5.2, includes cases where the reader is expected to imagine a space formed in a closed book through a succession of flat parts on adjacent pages. Here the coincidence of things on the left and right halves of an opening closed up is part of the semantic content of the book itself.

The structure of the paper is as follows. In Section 2 we present the situation that has been the main motivation for this study. This involves the combination of archival collections, information extraction, and digital images which are needed as part of the process of extracting information from the original documents. Having given motivating examples, we use Section 3 to review related work including document understanding, text where spatial arrangement conveys meaning and the use of Qualitative Spatial Representation to describe these spatial arrangements. At the end of Section 3 we introduce the variety of material structure of physical documents but we leave more detailed exploration of this area until later. Section 4 presents detailed case studies showing how spatial arrangement and material structure are important for understanding documents and extracting information from them. The key examples use specific documents from The National Archives. In Section 5 we consider the role of folds, tunnels, slits, and other holes in material documents. The section also suggests how the formal theory of qualitative spatial relationships might be developed to meet some of the challenges that we have highlighted. The paper ends with overall conclusions and topics for future work.

2 Archival Motivation

The National Archives is the archive of the UK government and holds around 1000 years of history. Although the vast majority of the holdings are only available in their physical form, around 10 million documents have been digitised. With the advent of handwritten text recognition (HTR) technology, and large language models, it is becoming possible to transcribe these documents at scale for computational analysis. In addition to this, initiatives such as IIIF are encouraging and making it easier for institutions to publish digitised collections online and making their textual content searchable. The reality of working with archival documents is that they are highly variable. Not only in terms of the content and layout of the pages themselves, but how they are stored, bound, and digitised. A case in point is catalogue reference MT 9/920/C at The National Archives, described as "Titanic. MT 9. Marine correspondence including SOS telegram from the Titanic, 1912". This is available to download at `https://discovery.nationalarchives.gov.uk/details/r/C5907412` as a 5 part PDF file. In total the PDFs contain 618 pages consisting of digitised images of a physical, paper file. The first image is of a manilla folder, which is then followed by a cover sheet (Figure 3) which details the contents and a section for minutes, which appear to capture notes and directions for onward processing of the information contained in the file. On the front page the minutes take up the lower two thirds of the page, while on the back they are separated into two columns one labelled "minutes" while the right hand one is clearly labelled "Minutes are not to be written on this side". What follows are several lined sheets of paper, including an introductory note and lists of passengers and crew on the ship. These lists consist of multiple columns which are laid out across two pages - starting on the back of one sheet, and going across to the front of the next. Each two page table is preceded by a page detailing the role of the people named in the table (passengers, crew, stewards, restaurant staff etc.). The tables for crew and passengers contain different columns. The remainder of the images are a mix of the aforementioned minute sheets, tabular information, correspondence on letter headed paper, envelopes, telegrams, handwritten notes, typed interview transcripts, pages with text cut from a different page glued to them, newspaper clippings, typed reports, bound instruction booklets, maps, and a few more manilla envelopes.

This is just one item of millions which have been digitised and are in theory ripe for exploitation by digital humanists with computational tools. Recent experience of the Material Culture of Wills project, suggests the practice is not so easy. Probate copies of wills are held

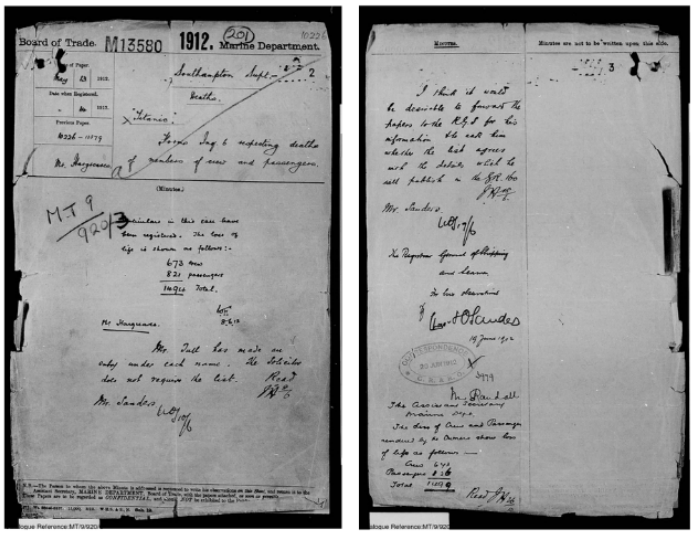


Figure 3 A sheet used for capturing minutes from the Titanic Correspondence file (MT 9/920/C).

in large bound volumes which were digitised from microfilm over 20 years ago. While the scribes wrote the wills in a continuous fashion, one will continuing immediately after the previous one, digital access is in the form of one PDF per will. The ordering of the wills in the catalogue is by date of probate not page ordering in the volume. Reconstructing the images of a volume of wills in their original order, with start and ends of each will marked and linked to its catalogue entry, was achieved through a combination of cryptographic hashes, text extraction, layout analysis, and manual intervention. It is clear from this example that digitisation was undertaken with providing access to individual wills in mind with no metadata retained to connect one page to the next, and that even in these structurally simple documents, layout has semantic meaning. If working with images of pages in bound volumes is a challenge at scale, what hope do we have with files such as the Titanic correspondence with all their variety?

It is not feasible to re-digitise millions of documents with much more detailed metadata about their material origin. Even if it were, it is less than obvious exactly what needs to be recorded and what structure might be appropriate. This raises the same issue as is found in working with the existing digitised versions. Experts in the documents may understand how different indentations, groupings of names, locations of signatures or marginal notes, for example, all play a part in extracting meaning. This knowledge, however, has to be communicated to computational experts who themselves need to describe the relevant patterns to process the data. What is needed is a language to express the spatial structure of the material in a way suitable for archival and other applications. This paper is a step towards such a language.

3 Related work

3.1 Document Understanding

The field of Document Layout Understanding is focused on identifying and interpreting structural aspects of documents. The DocLayNet dataset defines 11 distinct classes of structural element: namely caption, footnote, formula, list-item, page-footer, pageheader,

picture, section-header, table, text, and title. Layout recognition is also a key component of automated text recognition pipelines which translate segmented lines or words from an image into text. Another research area covering document layout is Document AI [19] which refers to the techniques for automatically reading, understanding, and analyzing business documents A survey of papers [22] published at ICDAR, the bi-annual International Conference for Document Analysis and Recognition, held since 1991, identified three main tasks being addressed by researchers: pre-processing for image quality, identification of regions with homogeneous content, and textual content recognition. Within the second category several sub-tasks were identified: text line segmentation, baseline detection, text detection, symbol detection. The authors further summarised the application of deep learning technology as mapping inputs (e.g., a digitised page) to outputs (page, region, table, text-line, bounding box, transcribed text, image patch, image), with the algorithms generally classifying at the pixel level. Wang et al. [30] build a layout graph where nodes are elements extracted from PDF metadata, and edges are created between nearest neighbours in each direction (left, right, up, down). A graph convolutional network is then trained to classify each element as one of the 11 classes of DocLayNet or the 5 of PubLayNet, another layout analysis dataset. The GraphDoc dataset [8] contains 80000 page images with associated layout graphs depicting spatial (left, right, above, below) and logical (parent, child, sequence, reference) relationships between between page elements (e.g. paragraphs, diagrams) found in scientific literature. Kim et al. [20] introduce the Document Understanding Transformer which extracts structured data from complex documents such as invoices and receipts for downstream tasks such as Information Extraction, and Visual Question Answering (where the task is to answer natural language questions about the content of an image). While relationships between elements have been used as features to classify elements, or documents, by type, to the best of our knowledge we have found no evidence of documents being represented by those relationships. Furthermore, the literature is focused on layouts of individual pages rather than multiple pages. Two data standards used in the humanities, the International Image Interoperability Framework (IIIF) and Text Encoding Initiative (TEI), are able to represent certain aspects of layout. The IIIF annotation standard stores bounding box coordinates of text on the page, and digitised images are held in a sequential order for viewing, but there is currently no concept of two sides of a sheet within the standard (although a working group has been set up to discuss this). TEI includes metadata tags for the cover of a book, front matter (such as contents), the main body, and back matter, and also has the concept of a group for material which are related in some way. It can also represent boundaries between pages and lines, chapters, paragraphs, and many of the concepts we may use to describe documents and the text the contain. It does not capture spatial relationships between elements on a page though.

3.2 Spatial Text

Text, whether handwritten, typed, or printed, flows in a direction dictated by the language; left to right (western), right to left (arabic), down the page (chinese). It also has orientation on the page, and from any particular perspective could be the right way around, upside down, at an angle, or even backwards (da Vinci famously wrote with a mirror). How much text fits on a page is dictated by the size of the page. A letter may cover both sides of a single sheet of paper, tabular information may continue on every page of a ledger, or its columns may span each double page spread. We may logically wish to think of a piece of text occupying a single page, even if physically it spans multiple sheets. Equally the text on the reverse side of a leaf may be contextually related but not a continuation of the text on the front, or it may

Figure 4 Line structure from Atkin's poem *Cannulation*. Image source [29, p23].

be completely unrelated beyond sharing the same piece of material. There is also a temporal element to text. A typed document may be heavily annotated with handwritten comments and corrections. It was common in centuries past for replies to letters to be written between the lines of the original, or even perpendicular to them, to save paper (cross-written letters).

Written language is read not only from screens but also from physical media such as paper. In much communication, the layout of the words and other building blocks of language is a significant part of the meaning. A particularly obvious example is concrete poetry [5], where the placement of the words on the page can be as important as choice of the words themselves. Poetry in which words are arranged to form pictorial images has a long history [17, 16, 5]. Well known examples include George Herbert's Easter Wings, poems in the form of altars, and the arrangement of words to imitate the shape of the tail of a mouse in *Alice in Wonderland* by Lewis Carroll.

Beyond poetry with a specifically pictorial function, the layout of poetry is often deliberate to connect with ideas in the wording. The lines in *Cannulation* by Polly Atkin [3] include the words:

`the self made small, importing little one dot iot iota drip.`

but the layout presents this in lines 16 to 20 of the text of the poem (see Figure 4) as follows:

```
the self made small, importing little
                                  one dot
                                           iot
                                                    iota
                                                              drip.
```

Here the succession of short lines progressing to the right relates to the process of successive dripping or measured puncturing. This layout is part of the poem itself in as fundamental a way as the choice of words themselves.

This kind of example concerns spatial relationships on a single page. The use of space by the author, and what matters to the reader or viewer, is not the exact coordinate location of words, but qualitative spatial relationships. In this context such relationships include the way that two words on successive lines appear with the later one further to the right than the earlier one but with a clear space vertically between the end of the first and the start of the second. For this spatial situation Allen Interval Relations have been found useful, as we explain next.

3.3 Allen Interval Relations

It has been suggested [29] that the Allen interval relations [2] can be used to record spatial relationships between lines of text. This is a different application area from the original context of intervals of time described by Allen. The Allen relations can be situated within a larger body of work on qualitative spatial representation. This area is concerned with general relationships such as "next to", or "inside", or "touching only at the borders". These relationships have had multiple applications to different kinds of spatial information, as explained in the survey [9]. The relevance of some of this work to the Digital Humanities is discussed in [29].

When representing layout with digital methods the position of text on the page is usually expressed in terms of pixel coordinates which may describe a rectangular bounding box, or a complex polygon, enclosing the text. While it may initially seem more appropriate to use a two dimensional QSR for rectangles, in practice most neighbouring bounding boxes will be disconnected - generally an element of text occupies its own discrete space. Bounding boxes can instead be represented as a pair of Allen intervals covering the vertical and horizontal extremities. Any element may be a part of multiple intervals: the horizontal span of a glyph will be part of a word, which is part of a line. Vertically the line is part of a paragraph, which is part of a page. We have a preference for the Extended version of Allen Intervals proposed by Renz [27], as we are able to express degree of overlap, although often only the degree from the perspective of one interval is important rather than both. In Figure 5, interval A overlaps both B and C in standard Allen terminology. Under the Extended algebra, the most part of A overlaps with both B and C, but while the most part of B also overlaps with A, only the least part of C overlaps with A, or mol(A,C) using the Extended Allen notation. The situation will dictate which is important so we separate degree of overlap for each interval from the Allen relationship. In this case we may have mo?(A,C), which indicates that mol and mom (Most Part overlaps Most Part) are equivalent.

Figure 5 Degree of overlap between intervals.

Other work using extended Allen relations can be seen in [6]. Unlike the present paper, the work described there is focussed on relationships between elements on a single page. Here we are dealing with the more complex case of relationships that may hold between entities on different pages.

3.4 Physical Documents

The physical structure of documents, whether handwritten manuscripts or printed books, or more generally forms such as scrolls, rolls, etc, is already a highly developed area of study. For example, see [15] for just one aspect of printed books, and [23] for the picture-scroll in China. However, the interaction of physical structure and information extraction, especially in combination with the need to process information computationally using existing images of the documents appears to have received relatively little attention. This situation is our primary motivation, as explained in Section 2.

The complexities of elaborate documents with multiple folding parts, loosely inserted pieces, pockets containing smaller documents, pages with holes, creases, etc are too extensive to address in full detail in this paper. It is even more challenging when such documents have been digitised in the past in ways that mean it is unclear how to reconstruct the original physical arrangement from a collection of digital images. We are able to illustrate some of the issues in a detailed case study of archival material in Section 4. Some of the challenges that occur with folds, cuts and holes are discussed in the subsequent section. We also use Section 5 to present a proposal for how formal theories of qualitative spatial relations might be developed to support the representation of relationships that occur within a space that is not rigid but can be modified, for example by turning the pages of a book.

4 Case Studies

Following analysis of a selection of digitised documents from TNA's online catalogue, and discussions with records experts, we have identified several common patterns for multi-page documents, in terms of both content and materiality, which we will model using QSR. We have not included rolled volumes in our analysis, as they are not turnable pages. Some of these patterns can be seen in Figure 6.

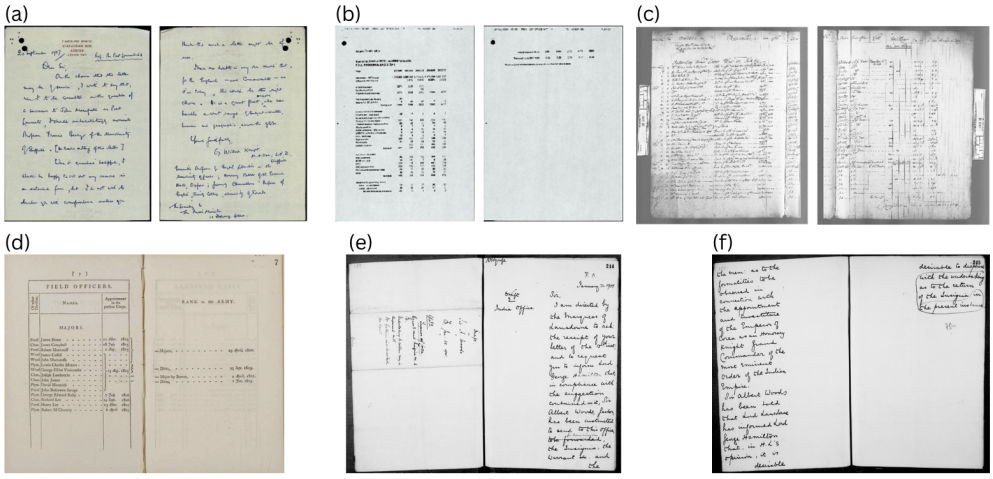


Figure 6 Selected double page layouts: (a) letter on two sides of sheet; (b) table continuing onto next sheet; (c) table with columns across two pages; (d) complex tabular information across two sheets; (e) open notebook pages, descriptive label at right angles, followed by text aligned to right and left hand margin note; (f) open notebook pages, opposite alignments leaving central margins.

Content patterns:

- Text, continues to next page (Figure 6(a))
- Text, related on next page but not continuous
- Tabular information, continuous over page (Figure 6(b))
- Tabular information, continuous across pages (Figure 6(c))
- Formatted text (e.g. indented), continuous over page
- Re-used material (e.g. reply to letter in line)

Material patterns:

- Single sheet
- Single sheet, folded
- Multiple sheets bound together (e.g. by treasury tag, or folder)

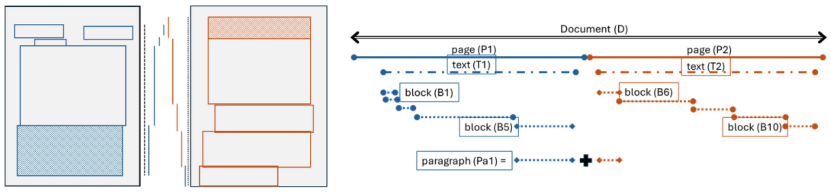


Figure 7 Left: A letter covering two sides of the same sheet, represented by bounding boxes. The vertical sides of the boxes form a set of intervals; Right: Allen interval representation of the document, pages, text, and text blocks.

- Bound book, horizontal open (left-right, right-left)
- Bound book, vertical open (bottom-top, top-bottom)
- Bound volume, contents originally individual sheets

4.1 Single sheet, continuous text

The document in Figure 6(a) is a letter written on two sides of a sheet of paper. We can model this document with bounding boxes around discrete blocks of handwritten text (e.g. the date, each paragraph, the signature). Between the two page images we have represented the vertical edges of the bounding boxes as a set of Allen intervals. The dashed and dotted lines represent the text on each page.

Figure 7 visualises Allen intervals for the entire document, each page, the entire text of each page, and each block of text identified by the bounding boxes. Additionally we create a new interval (Pa1) which represents a concatenation of the two shaded blocks in Figure 7.

The following example relations can be described between these intervals:

- startedBy(D, P1);
- finishes(P2, D);
- contains(P1, T1);
- contains(P2, T2);
- startedBy(T1, B1);
- finishes(T1, B5);
- finishes(T2, B10);

Physically the first handwritten line on page 1 begins just below the two holes in the sheet, but logically it is the beginning of the letter. So we could logically state startedBy(P1, T1) instead of contains(P1, T1). Blocks B5 and B6 are physically separate on different sides of the page, one at the bottom, one at the top. In terms of the text on the page they are a single paragraph, only broken up because of limitations dictated by the size of the paper. Logically, B5 and B6 make a paragraph, Pa1, which overlaps both pages, giving us overlaps(Pa1, P1) and overlaps(Pa1, P2).

4.2 Continuing indentation

Figure 8 shows indented text covering two pages. Each section, marked by bounding boxes, has multiple levels of indentation. The content of the final block of page 1 continues over to the first block of page 2. There is continuity in the indentation so the text on page 2 starts justified to the right, and then unindents after the block ends. Physically the indentation is

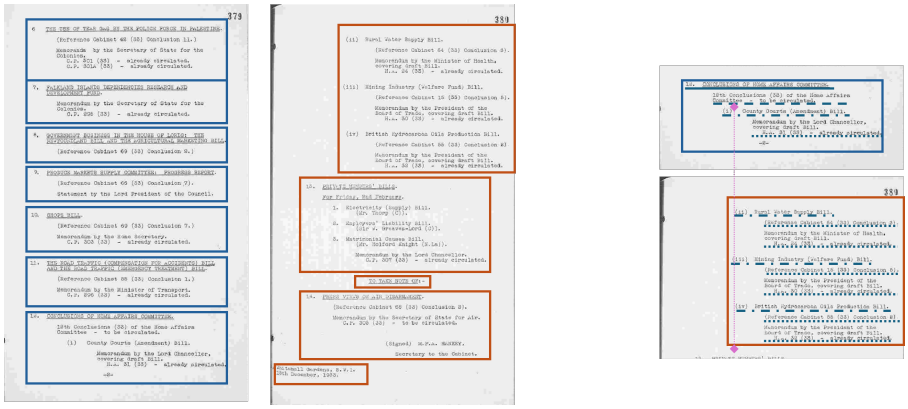


Figure 8 A document where indentation continues over the page. The indentations do not accurately line up, as indicated by the vertical dotted line.

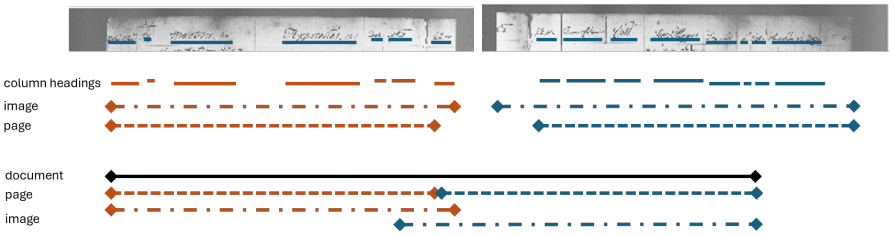


Figure 9 Column headings of a table which go across two pages.

not consistent between the two pages, with items (ii-iv) starting noticeably further to the right than item (i), as seen in the right hand image, but logically the four items all start together.

4.3 Tabular layouts

Figure 9 shows the columns headings of a table which spans across two pages. This representation shows the need for the digital as well as the physical as there is some overlap between the images with the first column on the second page appearing in both images.

A more complex tabular form is depicted in Figure 6(d). On the left hand page is a three column table. The table is a list of people, falling under the category of FIELD OFFICERS – a table which continues over several double pages of the book – and a sub-category of MAJORS. Names are listed in the order in which they were appointed to the rank, and are grouped by the use of brackets when more than one person was appointed on the same day. They are further grouped according to what appears on the right hand page. The title is RANK IN THE ARMY, which does not always appear on the right hand page when there is an entry. When the right hand page is completely blank it has been digitised separately to the left hand page. The entries in the right hand page are aligned with the centre of the outermost brackets on the left hand page. In addition, we note the word Ditto in some entries which means to take the same value as the entry immediately above. In the first case of Ditto, there is a noticeable space between it and the entry above. As a final complication, the sections of the book are catalogued by rank so Majors appear in a different PDF to the Captains, and so on.

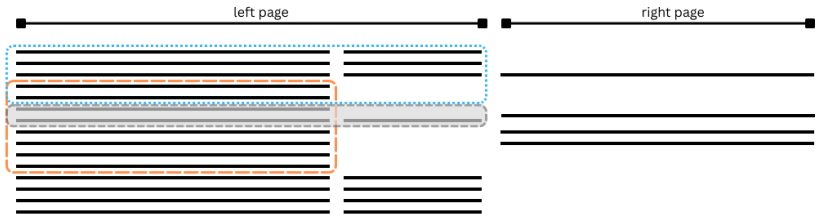


Figure 10 Complex tabular information where entries which share attributes (e.g. a date) are bracketed together. The text on the right hand page is not in the table but is aligned with its entries.

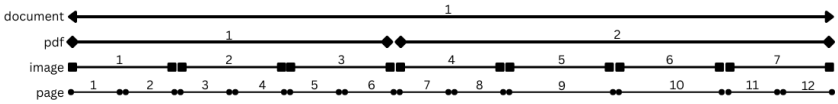


Figure 11 Multi-PDF document represented as intervals.

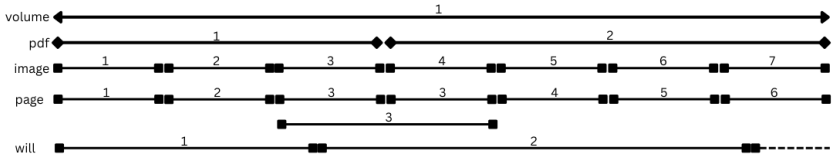


Figure 12 Volume of wills in PDFs represented as intervals.

The two pages are depicted as grouped intervals in Figure 10, with the groupings inferred from the bracketing represented by boxes around the appropriate intervals. We see that an interval may be a member of more than one group, with the consequence of group membership being dictated by the spatial relationships with other items on the same page, or the next one. The document itself is modelled as intervals in Figure 11 to illustrate the complex relationships between the physical and digital forms. Most images contain a double open page, while some show only one. The equivalent diagram for the wills mentioned in the introduction is shown in Figure 12. Extra complexity is introduced by the duplication of images between consecutive PDFs. We have both Finishes(Page3, PDF1) and Starts(Page3, PDF1), but Finishes(Image3, PDF1) and Starts(Image4, PDF2). The second Will overlaps both PDFs and images 3-6 but neither starts nor finishes any.

To represent the relationship of one will to the pages of the original book there is another technique besides giving an Allen relations between the will and each one of the pages. As each will is a continuous section of the book, without any gaps or interruptions, the relationship can be represented by a single Allen relation between the will and the sequence of pages in which the will appears. For example, in Figure 12 Will2 is found in pages 3–6. Thinking of this sequence of pages as a single interval denoted Extent2, the relationship Contains(Extent2, Will2) holds.

As well as capturing the logical flow of one page to the next we must also represent the physical relationships between pages to enable sense-making in the digital form. When documents are digitised they may be captured as double or single pages. When sifting

through digital images it is not always clear whether two images are of front and back of the same sheet, or whether they could be viewed side by side in their physical form. The pages of reporter style notebooks are flipped upwards, depending on the preferences of the owner, the writing on the upper page may be upside down or the same way up as the lower page when photographed. We have also encountered instances of standard opening notebooks where the book has been rotated through 90 degrees for digitisation due to the orientation of the writing. A page may therefore be represented by three orientations: natural reading, physical, and digitised. However, the orientation of the text to the page may vary in some documents. Qualitative Direction Calculi [26] such as the STAR algebra can represent relative orientation of a page to a viewer but further reasoning is required to derive the optimal orientation for a specific purpose. These details matter both to the person viewing search results in the image, and to an automated text recognition process. Another common pattern we see is for a space to be left on the side closest to the fold in the centre of a book when writing, as seen in Figure 6 e-f. This space acts as a margin, although it is not specifically marked as such. Any text in this space should be treated as marginalia, additional but related to the main text. The spatial relationship between text and marginalia alternates by page, depending if it is on the left or right on the fold.

5 Material Structures and Relationships

On a single page there are relationships between lines, between blocks of text, between the margins and blocks of text, etc. In a document with multiple pages all of those relationships are important on each page, but in addition there are relationships between the two sides of one leaf, as well as between leaves that share a common edge. There are relationships that depend on the viewer and on temporary configurations: two surfaces may be in contact when a book is closed and consequently be simultaneously visible when the book is open. However, once our books include pieces that fold out, two surfaces that can be seen simultaneously when the book is open need not be in contact when the book is closed. The complexity of the topic precludes a comprehensive account at this stage, but we draw the reader's attention to several ways that physical structure can impinge on the interpretation of information.

5.1 Folding

Consider maps that fold up. All of the sections on one side of the map are simultaneously visible when that side is uppermost to the viewer, but only a limited number of the sections would be in contact when the map is folded up. In the case of a folding map there are usually both horizontal and vertical folds and associated adjacency relations between sections in both dimensions.

There is a large literature on aspects of folding, both historical [14] and for practical applications in computational origami. However, most of this work is in quite a different direction to ours. For example, Justin [18] discusses questions such as the number of different ways a sheet such as a folding map can be folded up, or in a simpler case, the number of ways to fold up a sequence of panels arranged in a linear sequence when unfolded. It is notable that folding is of importance in archives in several ways such as the use of folding as a security device. This happens where a complex sequence of folds was used that would be almost impossible to restore if a document were intercepted and unfolded en route.

Consider Figure 13, which shows diagrammatically a book with two leaves (1 and 2 in the diagram) and a fold-out made up of three parts (3, 4, 5 in the diagram). Unlike the considerations in [18] about the combinatorics of folding up the 3-panel leaf 3-4-5, we

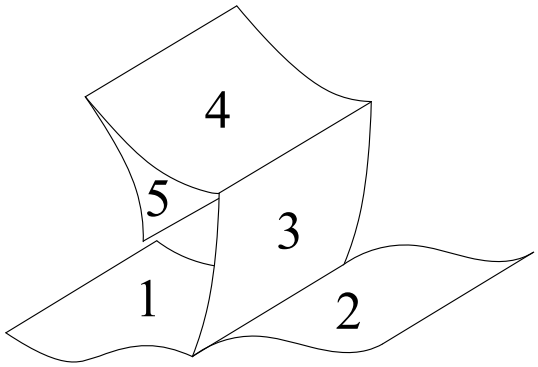


Figure 13 Diagram of part of a book with a fold out in three panels.

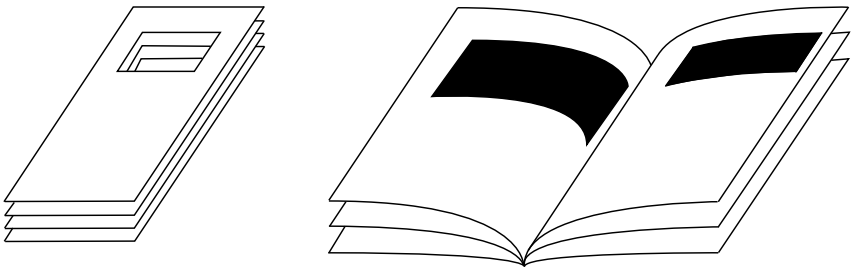


Figure 14 Books exhibit spatial relationships beyond those on a single page. A succession of pages with holes produces a tunnel when stacked together; an opening seen by a reader as opposing pages becomes two touching surfaces when closed. The reader can imagine tunnels without physical holes when regions on touching surfaces match up through several successive leaves.

generally assume a given way to fold it (which may be determined by the design of the book or historical precedence). One way in which we diverge from most studies of folding is that we are concerned with not just the material leaves and panels, but also with what may be present on the surfaces. Another is that qualitative relations have not been explored in this setting, whereas computational origami is generally concerned with geometric precision and complex three-dimensional forms.

5.2 Conceptual Holes

In *House of Leaves* by Michael Danielewski [11] the spatial structure of a book is essential to the content. Blocks of text in different typefaces are used in such a way that the reader is not following a one-dimensional progress through a sequence of words, but is required to navigate a network in which parts of one page link to parts of another. In places the text in a part of one page corresponds on the opposite side of the same piece of paper bearing the following page, to the same text repeated but transformed in a reflection. This provides the effect of the text on the second page functioning as the view of the text on the first page but seen from the opposite side, see Figure 15. This example of a relationship across two successive pages on a single leaf is extended further in the book as it forms part of a sequence of square blocks of text over several pages. This sequence, when the book is closed, makes up a tunnel through the leaves. The spatial structure of the tunnel is related to part of the narrative involving a tunnel itself. Here, words relate spatially not just to words on the same page, but to words elsewhere in a three-dimensional stack of pages.

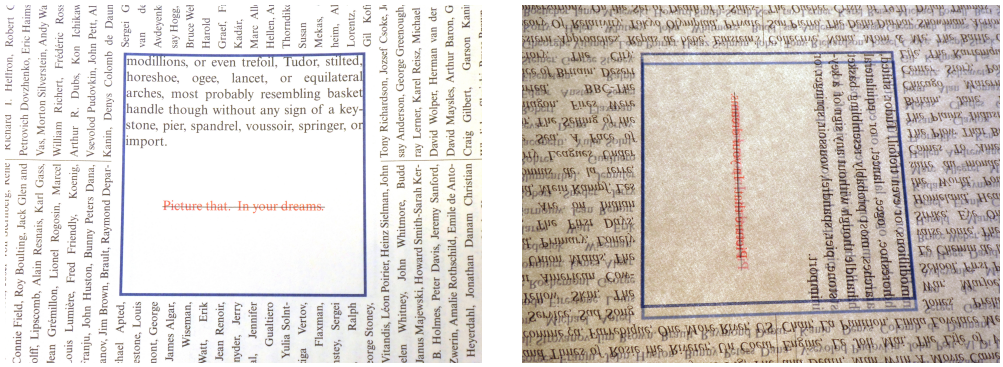


Figure 15 In *House of Leaves* [11] a sequence of pages have the two sides printed with what you would see if the paper were transparent. Although the practicality of printing this exactly has been challenging, the spatial relationship which is significant for the reader is the intended one rather than its physical realisation.

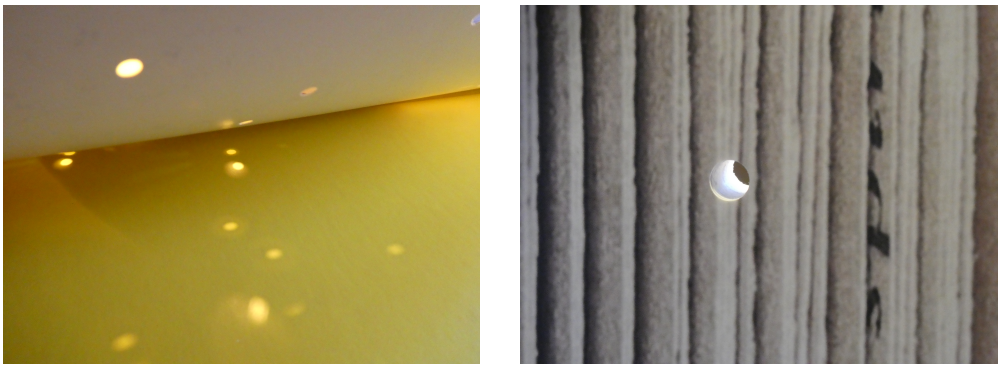


Figure 16 Unlike the tunnel in House of Leaves, these are physical holes. Examples from the dust jacket of *Holes* [7] (left), and *Inscription. Journal of Material Text* (right).

5.3 Cuts and Physical Holes

Unlike the tunnel that the reader imagines just discussed, books and documents can contain actual holes in their structure as shown in Figure 16. In other examples, a slit in a sheet allows a document to be folded up and read as a sequence of pages in a way that would otherwise be impossible. An example is shown in Figure 17, where the document is both a flat A3 poster on one side when unfolded, and an eight page zine on the other side when folded up. This type of slit in a sheet also appears in practical geographical examples in the map foldings patented by Falk [25].

5.4 Towards a formal theory

Documents involve a multitude of entities such as pages, leaves, openings, fold-outs, etc which make up a space where other entities such as individual lines of text, blocks of text, page numbers, illustrations, etc can be located. The categories of located entities within documents can be extended to encompass holes, creases, tears, edges of pages, and even missing pages. All these entities have relationships between them, for example two pages may constitute a single leaf or be opposite each other when a book is open. It may be possible to simultaneously view a page and some matter that folds out. Two blocks of text

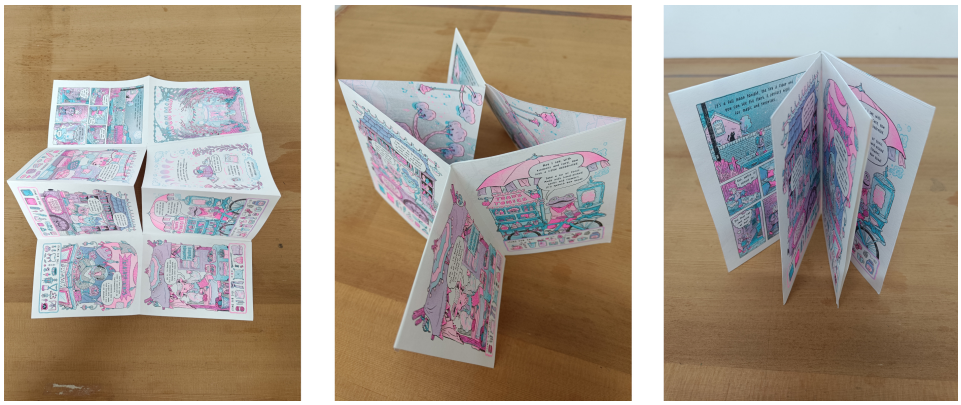


Figure 17 A single sheet with a slit folded into an 8-page zine: *MOON MART*, Kristyna Baczynski. More elaborate foldings with multiple slits are used for maps, one example being the Falk maps.

on different pages might be in contact once a book is closed. A comprehensive ontology giving a systematic account of all possible entities and relationships in this area would be a challenging project well beyond the scope of this paper. To illustrate some of the issues and to suggest a possible form for a more general study we consider a material space made up of four pages with a single fold. Imagine a greetings card for example. See Figure 18 which is explained below

The challenge for recording relationships between entities located within any material document space is that, unlike many situations with qualitative relationships, the configuration of the space can change and also may need to be described with respect to an observer. These are two different issues. For example, a book may be open at a particular page and turning the page changes the configuration in a way that can affect what surfaces are in contact. On the other hand, in a single sheet of paper, a hole might be on the left or on the right depending on the observer if there is no convention about what left and right mean. Both changes to configuration and changes to observer position will affect qualitative relationships.

To explore a specific case, consider the 4-page document in Figure 18. To make the example concrete, we imagine the document lying flat on a table in front of the reader. If the document is open and showing pages 2 and 3, as on the left of the figure, there are four possible ways of turning a page that bring us to a new configuration. The conventional action of turning a page to move forwards in a book is labelled in the figure as "turn to left". The turn in the opposite direction is labelled "turn to right". In such a simple document we can also imagine picking it up and turning a page underneath the document so the edge of the turned page moves away from the viewer and not towards them.

In addition to these four kinds of turning, which change the configuration of the document, there are also geometrical transformations in which the document is manipulated as a whole. These all come down to flipping about one axis or another (the line of the fold or orthogonally to it). Rotating by half a turn is the result of flipping about one axis and then about the other. Figure 18 shows some of these transitions between 12 different configurations and positionings of the document. Only some transitions are shown for clarity. The vertical lines with very short dashes all show rotations through half a turn. The diagonal broken lines all show flipping the document about the fold. The solid lines show turns (labelled t) and under turns (labelled u).

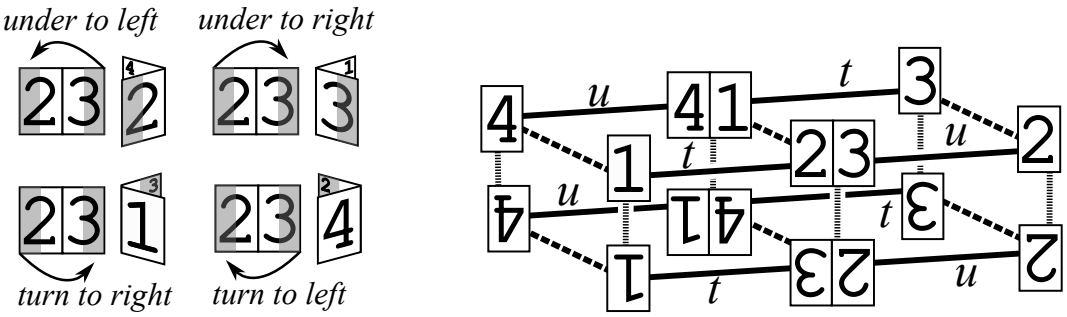


Figure 18 Space of a four-page document.

Now consider how qualitative spatial relationships are affected by some of the transitions between states shown in Figure 18. Look at the two grey regions shown occupying parts of pages 2 and 3 in the left of the figure. We have seen examples where the alignment of regions when a book is closed is significant. We have also seen in Figure 2 examples where paper is used that allows printing on one side to be read on the other side. It might thus be necessary to represent that the grey regions are actually in contact in some configurations, whereas in other ways of folding the document they can be "on top of each other" but not actually touching. When the document is unfolded and laid flat, they exhibit a third relationship.

The situation here appears different from regions which move and change relationships in examples such as two lakes becoming larger and merging together. Here, although the regions can move with the whole document, it is the space that they occupy that changes. To develop a formal theory that can accommodate this kind of change seems to need a way of representing how the relations in an appropriate calculus change along the transitions between different configurations. This suggests modelling the interaction between transitions and spatial relationships by having the space of configurations as a structure which acts (a term used here in its algebraic sense) on the algebra of relations; each transition being associated to a way of transforming the relations among themselves.

6 Conclusion

We have demonstrated that archival collections are very complex and have proposed that QSRs such as Allen Relations provide a logical framework for representing the relationships between the pages in documents, and the text which appears on them. A logical representation leads inevitably to the ability to reason over the documents based on the arrangement and layouts. The challenge is one of interpretation. While we can use qualitative representations to reason that one page follows the previous one, the examples we have presented require some level of knowledge to understand. We know intuitively if columns on one page share headings, and meaning, with those on the previous page. That the indentation of text is continuing from a previous page, and that indentation has an organisational purpose, implying parent child relations between blocks of text. The archivist knows about how paper was reused in times past and replies to letters were written at some angle – 45, 90, 180 degrees – to the original. Somehow we need a translation between the expert view of a document and the qualitative representation. Large language models certainly now provide an interface between human expression and computer interpretable language. Perhaps the framework and terminology of TEI is a starting point, which when combined with derived spatial relations, and a language model, could provide us with the tools to begin to work with large scale digitised collections such as we find in archives.

References

1 Ernest W. Adams. *Surfaces and Superposition*. CSLI Publications, Stanford, California, 2001.

2 J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843, 1983. doi:10.1145/182.358434.

3 P. Atkin. *Basic Nest Architecture*. Seren, 2017.

4 Fia Backström. *COOP a-script*. Primary Information, Brooklyn, New York, 2016.

5 N. Barker. *Visible Voices*. Carcanet, 2016.

6 Mark Bell, Eirini Goudarouli, John Stell, and Jane Winters. *Qualitative Space in Digital Diplomatics*, pages 175–196. Bölau Verlag, Vienna, 2025. doi:10.7788/9783412530921.175.

7 Roberto Casati and Achille Varzi. *Holes and other superficialities*. MIT Press, Cambridge, Massachusetts, 1994.

8 Yufan Chen, Ruiping Liu, Junwei Zheng, Di Wen, Kunyu Peng, Jiaming Zhang, and Rainer Stiefelhagen. Graph-based document structure analysis. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL: https://openreview.net/forum?id=Fu0aggezN9.

9 A. G. Cohn and J. Renz. Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of knowledge representation*, pages 551–596. Elsevier, 2008. doi:10.1016/S1574-6526(07)03013-1.

10 Suzanne Cotter, editor. *Julie Mehretu: A Transcore of the Radical Imaginatory*. Museum of Contemporary Art, Australia, Sydney, Australia, 2025.

11 Mark Z. Danielewski. *House of Leaves*. Pantheon Books, New York, the remastered full-color edition, 2000.

12 Johanna Drucker. *The Century of Artists' Books*. Granary books, New York City, 1995.

13 Dennis Duncan and Adam Smyth, editors. *Book Parts*. Oxford University Press, 2019.

14 Michael Friedman. *A History of Folding in Mathematics. Mathematizing the Margins*, volume 59 of *Science Networks. Historical Studies*. Birkhäuser, 2018.

15 P. Gaskell. *A new introduction to bibliography*. Oxford University Press, 1972.

16 S. Gross. The word turned image: Reading pattern poems. *Poetics Today*, 18:15–32, 1997.

17 J. Hollander. *Vision and Resonance. Two senses of poetic form*. Oxford University Press, 1975.

18 Jacques Justin. Aspects mathématiques du pliage de papier. In H. Huzita, editor, *Proceedings of the First International Meeting of Origami Science and Technology, Ferrara*, pages 263–277, 1989.

19 Sotirios Kastanas, Shaomu Tan, and Yi He. Document AI: A Comparative Study of Transformer-Based, Graph-Based Models, and Convolutional Neural Networks For Document Layout Analysis. *arXiv e-prints*, page arXiv:2308.15517, August 2023. doi:10.48550/arXiv.2308.15517.

20 Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517, Berlin, Heidelberg, 2022. Springer-Verlag. doi:10.1007/978-3-031-19815-1_29.

21 T. Y. Kong and A. Rosenfeld. Digital topology: a comparison of the graph based and topological appraoches. In G. M. Reed, A. W. Roscoe, and R. F. Watcher, editors, *Topology and Category Theory in Computer Science*. Clarendon Press, Oxford, 1991.

22 Francesco Lombardi and Simone Marinai. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10), 2020. doi:10.3390/jimaging6100110.

23 Shane McCausland. *The Art of the Chinese Picture-Scroll*. Reaktion Books, London, 2023.

24 Franz-Benjamin Mocnik and David Fairbairn. Maps telling stories? *The Cartographic Journal*, 55(1):36–57, 2018.

25 Mark Monmonier. Folding, unfolding. In *Patents and Cartographic Inventions: A New Perspective for Map History*, pages 105–135. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-51040-8_4.

26 Jochen Renz and Debasis Mitra. Qualitative direction calculi with arbitrary granularity. In Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap, editors, *PRICAI 2004: Trends in Artificial Intelligence*, pages 65–74, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. doi:10.1007/978-3-540-28633-2_9.

27 Jochen Renz, Reinhold Rauh, and Markus Knauff. Towards cognitive adequacy of topological spatial relations. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl Friedrich Wender, editors, *Spatial Cognition II, Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications*, Lecture Notes in Computer Science, pages 184–197. Springer, 2000. doi:10.1007/3-540-45460-8_14.

28 Barry Smith. Mereotopology: A theory of parts and boundaries. *Data and Knowledge Engineering*, 20(3):287–303, 1996. doi:10.1016/S0169-023X(96)00015-8.

29 J. G. Stell. Qualitative spatial representation in the humanties. *International Journal of Humanities and Arts Computing*, 13(1–2):2–27, 2019.

30 Jilin Wang, Michael Krumdick, Baojia Tong, Hamima Halim, Maxim Sokolov, Vadym Barda, Delphine Vendryes, and Chris Tanner. A graphical approach to document layout analysis. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi, editors, *Document Analysis and Recognition - ICDAR 2023*, pages 53–69, Cham, 2023. Springer Nature Switzerland. doi:10.1007/978-3-031-41734-4_4.

RCC-8 as a Benchmark for Diagrammatic Reasoning in Multimodal Foundation Models

Robert E Blackwell
School of Computer Science, University of Leeds, UK

Anthony G Cohn
School of Computer Science, University of Leeds, UK
Alan Turing Institute, London, UK

Abstract

Diagrams are commonly used to define relations in qualitative spatial calculi, and humans routinely rely on diagrams for spatial reasoning. Thus the question arises: can multimodal AI foundation models make use of diagrams for qualitative spatial reasoning? Using the Region Connection Calculus (RCC-8), a well-known qualitative spatial calculus, we investigate three capabilities: whether Vision–Language Models (VLMs) can recognise RCC-8 spatial relation diagrams, whether text-to-image and text-to-video models can illustrate RCC-8 relations, and whether Large Language Models (LLMs) can exploit diagrammatic reasoning. We show that, while no state-of-the-art model is fully reliable, the best models typically recognise vector diagrams more accurately than raster diagrams, albeit with problems recognising relations with a precise tangent. Diagrams with squares or circles are more readily recognisable than triangles or blobs. State-of-the-art LLMs can answer composition questions remarkably well, but there is little evidence that they benefit from explicit prompting for diagrammatic reasoning. Although state-of-the-art image and video generation models are widely used to, for example, produce social media content, they struggle to provide accurate RCC-8 relation or conceptual neighbourhood illustrations except when SVG is the output format.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning

Keywords and phrases Large Language Models, Foundation Models, Vision-Language Models, Spatial Reasoning, Diagrammatic Reasoning

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.4

Supplementary Material *Dataset (Dataset)*: `https://github.com/RobBlackwell/cosit2026` [5] archived at `swh:1:dir:d771bd22f1b7bc35e03e0659eb7f973ccc8fbafb`

Funding This work was supported by the Fundamental Research priority area of The Alan Turing Institute, by the Economic and Social Research Council (ESRC) under grant ES/W003473/1, by the EPSRC under grant EP/Z003512/1, and by the Special Funds of Tongji University for the "Sino-German Cooperation 2.0 Strategy".

Author Contributions REB & AGC jointly conceived the idea for the paper; REB conducted all the experiments and wrote the first draft of the paper; AGC & REB jointly edited subsequent drafts.

1 Introduction

"Many types of everyday and specialized reasoning depend on diagrams: we use maps to find our way, we draw graphs and sketches to communicate concepts and prove geometrical theorems, and we manipulate diagrams to explore new creative solutions to problems" [29]. Diagrams provide external representational support to cognitive processes, for example, diagrams avoid you having to memorise the map of a neighbourhood [29]. Diagrams make abstract properties and relations accessible, for example data visualisation allows us to see patterns in data [29].

© Robert E Blackwell and Anthony G Cohn;
licensed under Creative Commons License CC-BY 4.0

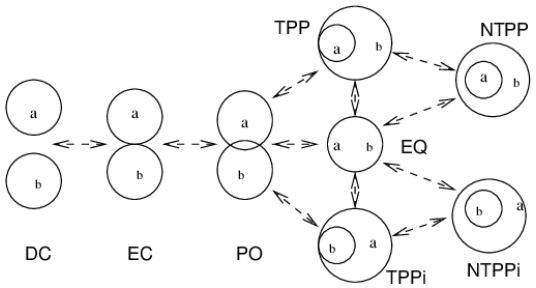


Figure 1 The eight relations of the RCC-8 calculus illustrated in 2D [11]: DC (Disconnected), EC (Externally Connected), PO (Partially Overlapping), TPP (Tangential Proper Part), NTPP (Nontangential Proper Part) and EQ (Equals); TPPi and NTPPi are the inverses of TPP and NTPP respectively since they are asymmetric.

Diagrams are often used to define relations in qualitative spatial and temporal calculi, for example the Region Connection Calculus, RCC-8 [25, 11]. RCC-8 is a qualitative spatial calculus for representing and reasoning about spatial relationships between non-empty regular closed regions of uniform dimension in a topological space. Figure 1 illustrates the RCC-8 relations and the RCC-8 conceptual neighbourhood (CN). The CN of a relation R between entities x and y is the set of all possible relations which might immediately be obtained next assuming x and/or y were to continuously deform or translate, without any other relation needing to hold in between. The composition table (CT) specifies the set of all possible relations $R_3(x, z)$ that can hold if $R_1(x, y)$ and $R_2(y, z)$.

Much progress has been made in the field of Foundation Models, with Large Language Models (LLMs) demonstrating remarkable reasoning capabilities [32]. Vision Language Models (VLMs) extend LLMs with vision encoders so that they can accept images as input in addition to text [22]. Generative image and generative video foundation models produce visual content (images and video) from textual descriptions [24, 26].

The question arises, can foundation models use diagrams for reasoning? In this paper we use RCC-8 to test the ability of foundation models to recognise relation diagrams, to reason with diagrams, and to illustrate relations diagrammatically.

2 Related Work

Diagrammatic reasoning has long been recognised as a fundamental component of human cognition. For example Bauer et al. [3] showed that diagrams are not merely illustrations but act as representations that structure information spatially, enabling people to reason faster and draw more valid conclusions than with verbal premises alone. Barker-Plummer [1] explored the role of diagrams in mathematical proofs and showed how diagrams can contribute to, rather than merely accompany, formal mathematical reasoning: “Far from being an expendable aid diagrams play an essential role in the communication of mathematical meaning”. Jamnik [19] developed systems that integrate diagrammatic and symbolic reasoning, showing how the two can complement each other within formal proofs rather than being treated as separate modes of inference.

In seminal work, Larkin and Simon [20] compare diagrammatic and sentential representations, and note that the former have a particular advantage that the former support what they call “perceptual inferences” which make it very easy for humans (at least) to extract information or knowledge which is much less easy to comprehend in a sentential representation. Diagrammatic representations can also simply search for particular kinds of information. Further analysis along these lines can be found in Glasgow et al. [9].

Xie et al. [31] tested LLMs with RCC-8-like relations between US states and time zones, finding that while LLMs achieve above-chance performance, their reasoning is unreliable. The ability of LLMs to reason about RCC-8 has been investigated [12, 16], however these analyses focused on model performance when presented with purely linguistic descriptions of spatial relations and queries. Although they identified systematic failure modes and partial abilities, these studies did not address multimodal input, generating diagrams, or reasoning that explicitly used diagrams.

Hsu et al. [17] proposed a framework that lets an LLM generate and interpret diagrams via a visual scratchpad, showing that this can improve reasoning over an inference-only LLM on certain text-based tasks. However, experiments showed that it still underperforms a single fine tuned LLM, suggesting that current vision foundation models struggle to understand diagrammatic abstractions. Other work that investigates LLMs' abilities to understand diagrams includes work on flowcharts [27, 34] and, using a specifically trained Multimodal LLM (MLLM), UML diagrams [2].

Several lines of work suggest that visualisation, or simulated visualisation, may enhance model reasoning. Wu et al. [30] report that "Vision of Thought" (VoT) prompting improves spatial reasoning performance in LLMs. The method prompts the LLM to "Visualize the state after each reasoning step". They report substantial improvements in tasks requiring spatial and geometric reasoning. Borazjanizadeh et al. [6] test large multi-modal models by asking the AI models to generate a conceptual diagram for a domain before solving a problem. Their results indicate that explicitly producing a diagram can guide subsequent reasoning, though the evaluation does not focus on formal spatial calculi.

Zhang et al. propose Diagram of Thought reasoning (DoT) in LLMs [35], where a problem is iteratively split into a kind of directed acyclic graph by LLMs fulfilling the roles of proposer, critic and summariser. Although the method uses the word "diagrammatic", the process has little to do with the kind of diagrammatic reasoning we are concerned with in this paper.

The ability of VLMs to understand spatial configurations in photographs has been investigated, e.g. [15, 10]. However our interest is in qualitative spatial reasoning rather than image interpretation *per se*. There has been some work on getting (M)LLMs to generate diagrams but usually from quite explicit textual instructions as to what the diagrammatic objects should be, and how they should be connected, for example DiagrammerGPT [33]; this contrasts with the task in the present paper which is more abstract.

3 Experimental Design

Based on Tylén's work [29], we define diagrammatic reasoning as the use of external pictorial representations (diagrams) to make abstract relations perceptible and, through their manipulation, to generate or extract information for problem-solving. Doing this requires the ability to generate diagrams, to recognise relations implicit in diagrams (particularly in our case spatial relations) and to be able to use these capabilities to improve the model's reasoning capabilities. These abilities inspire our choice of tasks in our experiments.

We use RCC-8 to generate questions that we pose to foundation models with textual input modalities, diagrams for models with image input modalities, and prompts for image and video generation models. In all experiments, we prefix the prompt with the following explanatory text:

```
RCC-8 is a qualitative spatial calculus for representing and reasoning about spatial
relationships between non-empty regular closed regions of uniform dimension in a
topological space.  It consists of eight jointly exhaustive and pairwise disjoint
binary spatial relations.  DC means that x and y are disconnected and do not have
intersecting boundaries.  EQ means that x and y are coincident.  PO means that x
and y have a region z in common but neither is part of the other.  TPP means that
the boundaries of x and y intersect, x and y are not coincident, and x is a part of
y.  EC means that x and y have intersecting boundaries but do not share any interior
parts.  NTPP means that the boundaries of x and y do not intersect, x and y are
not coincident, and x is a part of y.  TPPi means that the boundaries of x and y
intersect, x and y are not coincident, and y is a part of x.  NTPPi means that the
boundaries of x and y do not intersect, x and y are not coincident, and y is a part
of x.
```

Some models have hyper-parameters and configurable methods of sampling, including setting *temperature*, *top_p* and a *random seed*, whilst others do not. For simplicity, and to avoid favouring one model over another, we accept each model's defaults, unless otherwise stated. Similarly, "prompt engineering" is the practice of customising a prompt, often in a model-specific manner; we avoid prompt engineering as far as possible, preferring to keep the prompt consistent for each experiment and model. Each prompt was presented as a new, "zero-shot conversation" so that models retained no context from earlier prompts.

Where the output of a model is text (i.e. relation recognition and CT questions), we use automated evaluation to score the results (based on regular expressions and string matching). For images and videos, we score the results by hand (by one of the authors, with cross checking by another author), assessing the correctness of the generated media. In all cases the final answer must be strictly correct to score one, or zero otherwise. We evaluate LLM performance by calculating the total score for each experimental repeat and considering the mean and standard deviation (for overall model accuracy comparisons) and median (convenient in figures and tables).

3.1 Relation Diagram Recognition

We first test the ability of VLMs to recognise diagrams that depict the RCC-8 relations between two regions labelled x and y. We vary the shapes of x and y (circles, triangles, squares, and blobs[1], Figure 2). We also vary the foreground and the background colour – black on white, white on black, green on red (8% of European Caucasian men have red-green colour vision deficiency [4] and so it is reasonable to ask whether this deficiency is reflected in VLM performance), orange on blue (orange and blue is the colour combination that can most readily be discriminated by people with colour vision deficiency [23]) and two adjacent shades of red (red is the most difficult primary colour within which to distinguish shades). We use Scalable Vector Graphics (SVG)[2] standard colours for "black", "white", "red", and "green". Based on [23], we use #E69F00[3] for orange, and #0072B2 for blue. The additional shade of red is #FE0000. Eight relations with five colour formats and four shape formats gives 160 images. We present images in SVG format and then Portable Network Graphics (PNG)[4] format. We test a representative selection of VLMs – *OpenAI GPT-5-2* (henceforth *GPT-5.2*), *xAI Grok 4* (henceforth *Grok 4*), *Gemini 3 Flash Preview*, and *Qwen3-VL-235B-A22B-Instruct* (henceforth *Qwen3 VL*) – using the following prompt:

[1] In this paper, we define a blob to be any irregular curved shape with convexity and concavity.
[2] `https://www.w3.org/groups/wg/svg/`, retrieved February 2026
[3] Hexadecimal (hex) RGB colour notation represents colours using a 6-digit code preceded by a hash (#RRGGBB), where pairs of digits (00-FF) represent the intensity of Red, Green, and Blue, respectively.
[4] `https://www.w3.org/TR/png/`, retrieved February 2026

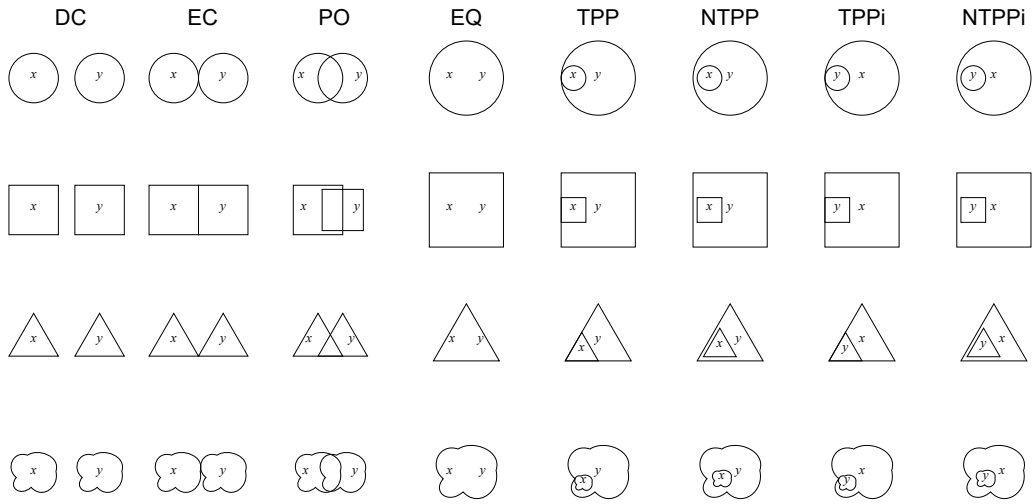


Figure 2 RCC-8 relation diagrams for circles, squares, triangles, and blobs. We use these diagrams in both SVG and PNG format to test the ability of VLMs to recognise the RCC-8 relations.

```
<RCC-8 Description>
Which of the eight relations best describes this image?  Answer with DC, EC, PO, EQ,
TPP, TPPi, NTPP, or NTPPi and no additional, extraneous text.
<IMAGE>
```

3.2 Relation Diagram Generation

For each relation $R \in \{\text{DC}, \text{EC}, \text{PO}, \text{EQ}, \text{TPP}, \text{TPPi}, \text{NTPP}, \text{NTPPi}\}$, we give the RCC-8 description and then ask image generation models to "Draw a diagram to illustrate the R(x,y) relation." The VLMs that we used for diagram recognition are not image generation models (they do not output raster graphics), and so we test *Google Gemini 3 Pro Image Preview*, *Google Gemini 2.5 Flash Image*, *OpenAI GPT Image 1*, *OpenAI GPT Image 1.5*, and *Z Image Turbo* (a 6B parameter model that runs locally, experimentally, with Ollama[5]). We also test *GPT-5.2* by adding "Ensure that the diagram is in SVG format." to the prompt (it is not an image generation model, but can generate SVG).

3.3 Compositional Reasoning

We next test the ability of models to perform compositional reasoning with RCC-8 relations. We assemble 64 questions from the RCC-8 CT (after [12]) of the form "If $R_1(x, y)$ and $R_2(y, z)$ what are all the possible relations that can hold between x and z?" We test the VLMs used in the earlier experiment as well as *Claude Sonnet 4.4*, and *Deepseek V3.2*. (When used with *reasoning-effort* set to *high*, *Deepseek V3.2* is one of the few state-of-the-art models to give a full reasoning trace.)

We first conduct a baseline experiment where we ask LLMs the questions without any specific instructions about the reasoning approach to be employed, for example:

[5] `https://ollama.com/blog/image-generation`, retrieved February 2026.

```
<RCC-8 Description>
I will now ask a question about these relations.  The answer may include more than
one possible relation.
List all possible relations as a comma separated list, with no extraneous text,
choosing only from the relations DC, EQ, PO, EC, TPP, TPPi, NTPP, and NTPPi.
If DC(x,y) and DC(y,z) what are all the possible relations that can hold between x
and z?
```

We then test Vision-of-Thought (VoT) prompting [30] by modifying the prompt to include the line "Visualize the state after each reasoning step and then list each possible relation". We also test whether explicitly prompting for diagrammatic reasoning improves LLM accuracy by including the instruction "Draw a labelled diagram to illustrate each possible relation and then list each possible relation."

3.4 Conceptual Neighbourhood Illustration

Videographic. We test whether a video generation model (OpenAI *Sora 2*[6]). can illustrate the RCC-8 CN. For all pairs of relations R1, R2 in the CN, we use a prompt consisting of the RCC-8 explanation text above and the instruction:

```
<RCC-8 Description>
The conceptual neighbourhood of a relation R between entities x and y is the set of
all possible relations which might immediately be obtained next assuming x and or y
were to continuously deform or translate, without any other relation needing to hold
in between.
For example, R1(x,y) is a conceptual neighbour of R2(x,y)
Illustrate the transition from R1(x,y) to R2(x,y) using a video with two regions
labelled x and y.  Start at R1(x,y), then go to R2(x,y) translating or scaling the
regions accordingly.
```

Frames. Finally, we use a similar prompt to the *videographic* experiment above, instead asking "Illustrate the transition from R1(x,y) to R2(x,y) using five image frames, as if from a video. Use two regions labelled x and y. Start at R1(x,y), then end at R2(x,y), translating or scaling the regions accordingly in the intermediate frames". This prompt allows us to test both image generation models (generating PNG) and LLMs (those which can generate SVG).

4 Results

4.1 Relation Diagram Recognition

None of the VLMs tested was able to recognise all the relation diagrams, but *GPT-5.2* and *Grok 4* correctly recognised all DC, PO, NTPP, and NTPPi diagrams when presented in SVG format (Figure 4). None of these four relations requires a precise tangent. *GPT-5.2* was able to reliably recognise EC when relations were presented in PNG format, but the confusion matrix in Figure 5 shows that *GPT-5.2* tends to overestimate EC so this result is not surprising.

The two best performing models (*GPT-5.2* and *Grok 4*) recognised more relation diagrams in SVG format than in PNG format (87% compared to 72%, and 83% compared to 53% respectively). None of the models was able to recognise all of the tangential relations TPP or TPPi. DC was the easiest relation to recognise and TPPi the hardest. The lack of symmetry is notable – i.e. one would expect that the inverse relations would be just as easy to recognise the regular relations and so it is unclear why the results for TPP and TPPi and NTPP and NTPPi are so different.

[6] `https://openai.com/index/sora-2/`, retrieved February 2026.

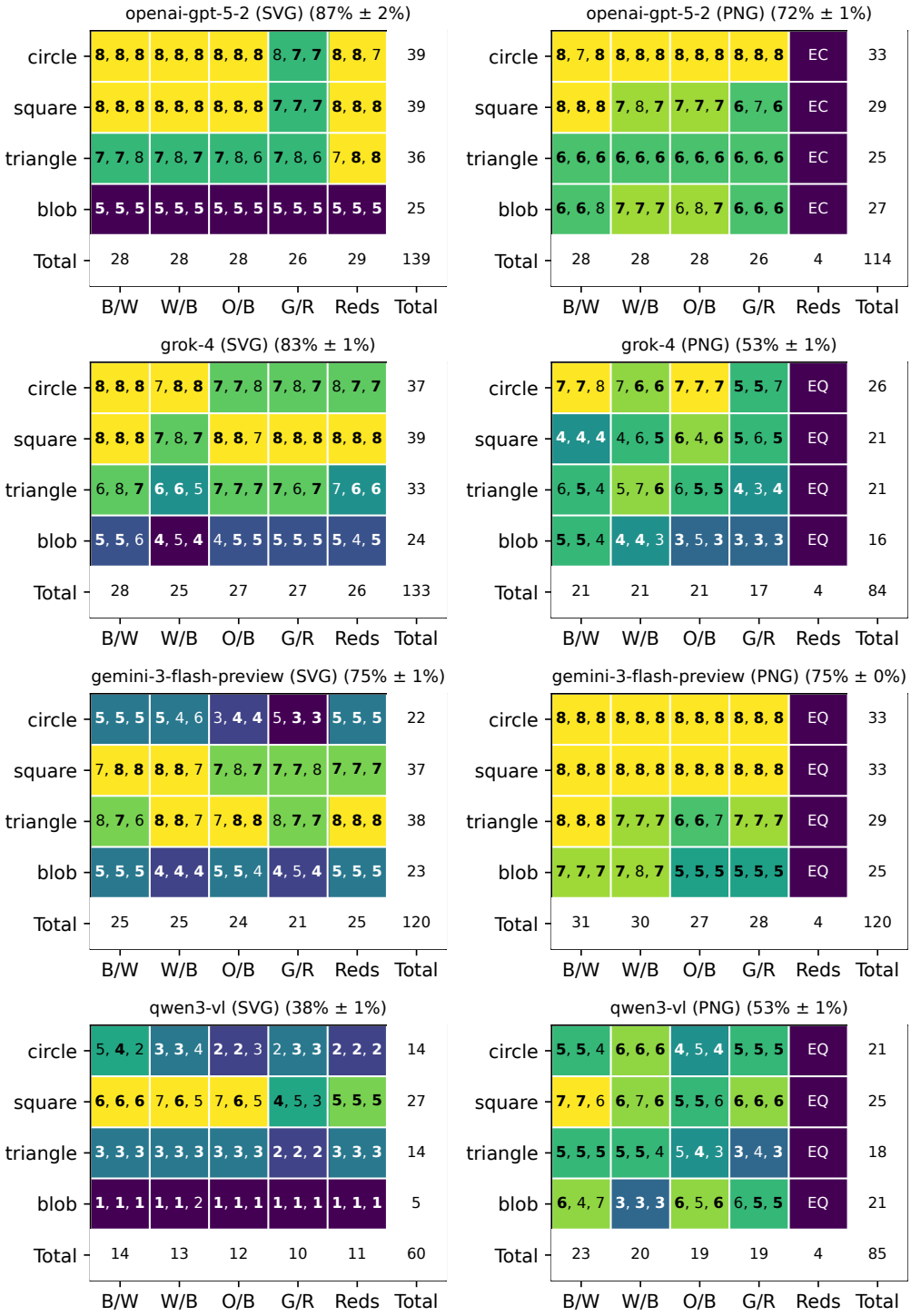


Figure 3 Vision-Language Model relation diagram recognition comparing shape and colour combinations for SVG and PNG by model. Each model run has 160 questions (eight relations × five colour combinations × four shapes), so the score in each cell is out of 8, with n=3 repeats. Rows and columns are ordered by the sum of the median scores, so shape and colour combinations with higher median performance appear first. The highest scoring cells in each column are in bold. Heat map colours reflect the median score. Where a cell contains 1,1,1 and the same relation was recognised across all repeats, we replace that cell with the relation name.

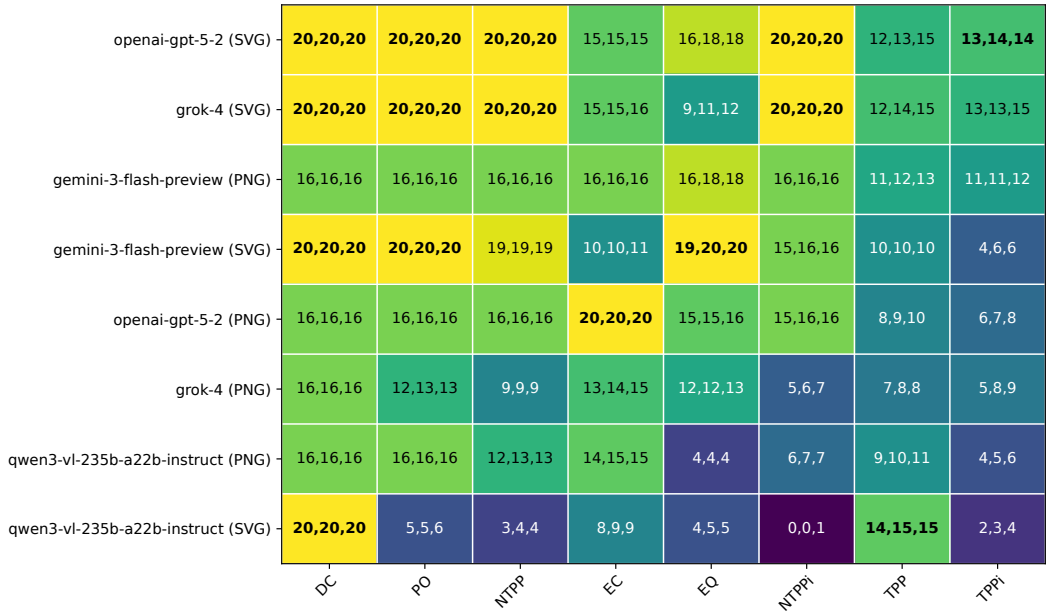


Figure 4 Vision-Language Model relation diagram recognition for each of the RCC-8 relations. Each model run has 160 questions (eight relations × five colour combinations × four shapes), and so each score is out of 20, with n=3 repeats. Rows and columns are ordered by the sum of the median scores, so models and relations with higher median performance appear first. The highest scoring cells in each column are in bold. Heat map colours reflect the median score.

When presented with SVG diagrams, *GPT-5.2* predicted NTPP when the answer was TPP (20/60), NTPPi when the answer was TPPi (19/60), and DC when the answer was EC (11/60), but never the opposite (Figure 5). On one occasion, *GPT-5.2* failed to give an answer to a PNG question at all, when the answer was NTPPi.

Overall, models tended to recognise more relation diagrams using squares or circles, than triangles, or blobs (Figure 3). E.g., comparing squares and blobs for *Grok 4*, a two-sided Welch's t-test shows that the difference is statistically significant ($t = 18.37$, $p = 0.0001$).

Models tended to recognise more relations in black and white than in colour (Figure 3). SVG diagram recognition was largely invariant to colour choice; this is not surprising since the colour is simply an attribute on a vector and the basic shape recognition would expected to be unaffected by varying colours. When coloured PNGs were used, orange on blue was most accurately recognised, followed by green on red, with two shades of red being the most challenging. The last result is perhaps not surprising – this task would be virtually impossible for all but the most perceptually acute humans, but the fact that the LLMs in general deteriorated in the same way as many humans from black/white, to orange/blue to green/red is more surprising. In the "reds" column of the PNG diagrams in Figure 3, it is not surprising that EQ is the only relation recognised by three of the models – the image would just appear to be single large red rectangle; why *GPT-5.2* interpreted this as EC is puzzling.

4.2 Relation Diagram Generation

All the image generation models tested except *Z Image Turbo* were able to generate plausibly correct diagrams of DC and PO but only *GPT-5.2* (not an explicit PNG image generation model, but prompted to output SVG) generated correct NTPP diagrams (Figure 6). *GPT-5.2*

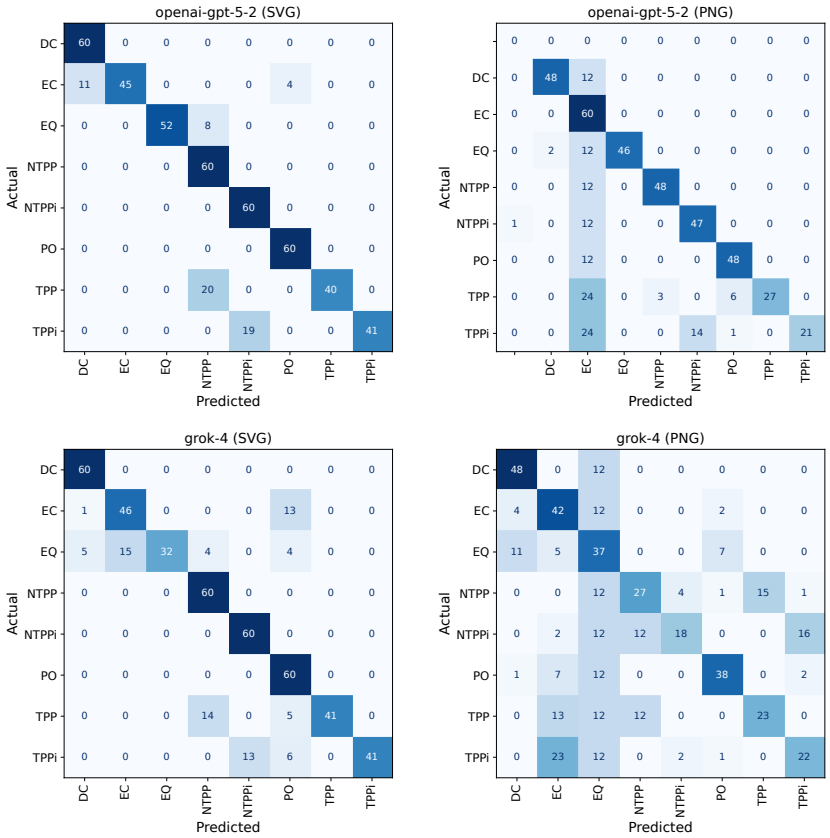


Figure 5 Relation diagram recognition confusion matrices for the best performing models (*GPT-5.2* and *Grok 4*), with SVG on the left and PNG on the right.

was the best model overall (7/8). *Gemini-3-pro-image-preview* was the best image generation model (5/8) and *Z Image Turbo* the worst (0/8, despite being able to generate photo-realistic images [8], *Z Image Turbo* produced diagrams that look more like maps instead of RCC-8 relations (not shown).

GPT Image 1 always generated black on white diagrams, *GPT Image 1.5* and *Gemini 2.5 Flash Image* always generated colour images. Overall, ten diagrams were black and white and eight orange and blue. 25 diagrams contained circles or ellipses and 16 contained blobs – no diagrams used triangles or squares.

4.3 Compositional Reasoning

For our best performing models (*GPT-5.2* and *Grok 4*), the choice of baseline, VoT, or diagrammatic reasoning prompting makes little difference (Figure 7). However for *Gemini 3 Flash Preview* and *Qwen3 VL*, diagrammatic reasoning shows marginal improvement (46 compared to 37 and 33 compared to 16 respectively). Both models also show improvement with VoT (47 compared to 37 and 23 compared to 16 respectively). *Claude Sonnet 4.4* and *Deepseek V3.2* (with default reasoning effort) showed some improvement with VoT (42 compared to 37 and 27 compared to 17 respectively), though neither show an improvement when using diagrammatic reasoning.

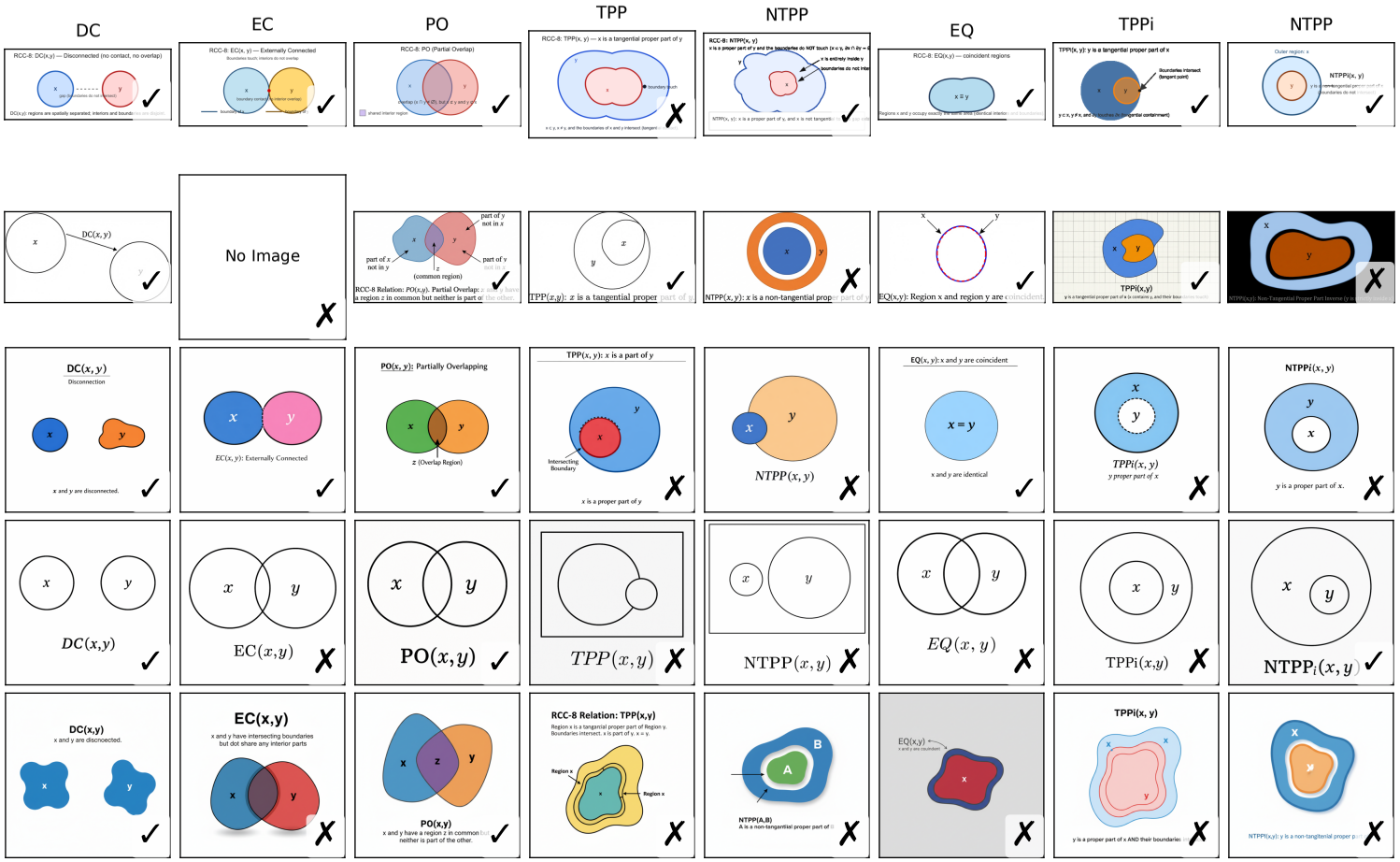


Figure 6 Relation diagram generation by model. gpt-5.2 (high) with SVG output (top, score 7), gemini-3-pro-image-preview (score=5), gpt-image-1-5 (score=4), gpt-image-1 (score=3), and gemini-2-5-flash-image (score=2). z-image-turbo scored 0 and is not shown. Ticks and crosses represented correct and incorrect answers respectively. This figure gives an overall impression of image generation results. Higher resolution images, with more readable text, are available in the GitHub repository that accompanies this paper.

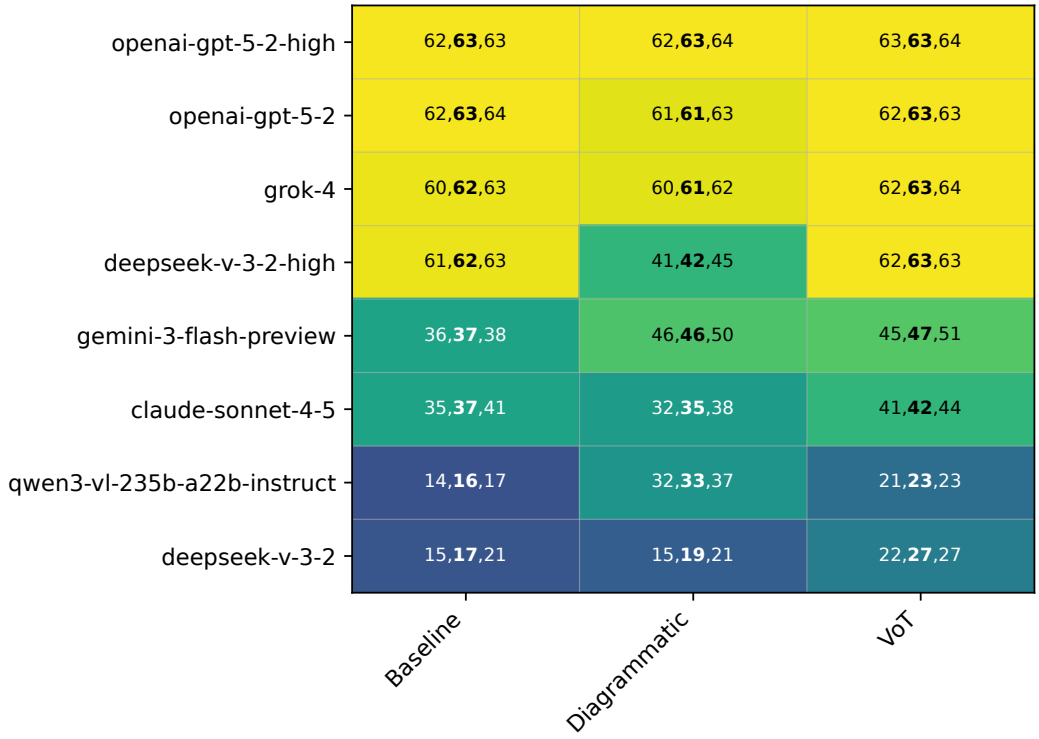


Figure 7 RCC-8 CT correct answers by model for baseline, diagrammatic reasoning and Vision of Thought experiments. Each model run has 64 questions (eight relations × eight relations), so each score is out of 64. n=3 repeats are shown with the median in bold. Rows are ordered by the sum of the maximum values across columns. Heat map colours reflect the median score.

When prompted for diagrammatic reasoning, models often mentioned or attempted ASCII art diagrams in their reasoning trace (*Deepseek V3.2* with high reasoning effort, 163/192) or reasoning summary (*GPT-5.2* 121/192 and *GPT-5.2* with high reasoning effort, 139/192). Neither model ever mentioned SVG, and neither model showed this behaviour with VoT prompting.

4.4 Conceptual Neighbourhood Illustration

Videographic. Using *Sora 2* with the default eight second video length, only 3/22 of the videos were plausible illustrations of RCC-8 CN neighbourhood transitions (TPP to PO, TPP to NTPP, and NTPP to TPP). Only 13/22 correctly depicted the starting relation in the first frame (Figure 8). 11/22 videos used blue and orange as the two main colours. 14/22 videos used circles or ellipses and only 6/22 used blobs. There were no squares, but one video used lozenge shapes (EQ to NTPPi). All videos had narration and this seemed to get cut off abruptly, suggesting that the videos were too short. However we also tried some 12 second videos (the longest available) and these exhibited the same problem.

Frames. *GPT-5.2* generated all 22/22 sets of five frames plausibly using SVG (see Figure 9 for an example). In some cases the SVGs were cropped, making assessment challenging. Blue and orange and blue and red were the most frequent colour choices (both 10/22) although shades varied. In all cases except one (21/22) circles were used to illustrate the regions, but

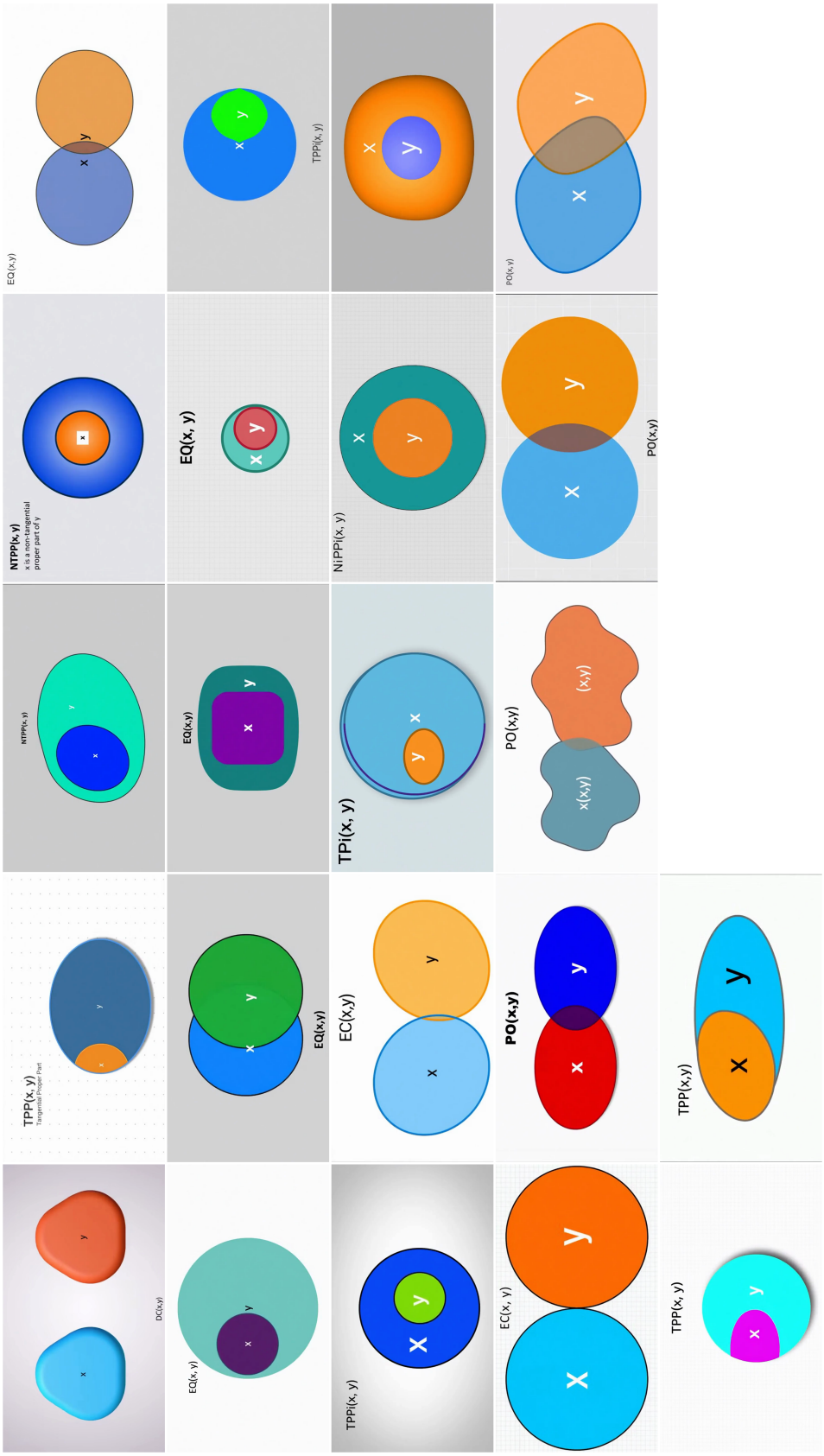


Figure 8 First frames from the *Sora 2* CN videos (n=1). This figure gives an overall impression of CN video generation results. The videos themselves are available in the GitHub repository that accompanies this paper.

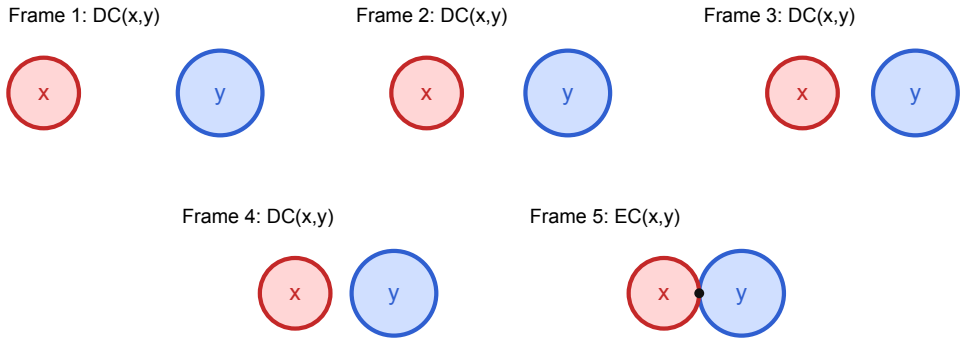


Figure 9 Frames generated by *GPT-5.2* for the CN transition from DC to EC (frame labels are part of the generated SVG). The generation also includes this commentary: Frame 1 – **DC(x,y)** (clearly separated) Frame 2 – **DC(x,y)** (closer, still a gap) Frame 3 – **DC(x,y)** (very close, still disconnected) Frame 4 – **DC(x,y)** (almost touching; still no boundary intersection) Frame 5 – **EC(x,y)** (boundaries touch at exactly one point; interiors disjoint) In Frames 1–4, **x** and **y** remain **DC** because there is still a nonzero gap between their boundaries. In Frame 5, the regions become **EC** because their **boundaries intersect (touch)** while their **interiors do not overlap**.

in one case (NTPP to TPP), squares were used – a sensible choice because the TPP relation shares a line for squares (assuming all boundaries are vertical or horizontal rather than, say at 45° in which case a point contact would be possible) but would only be a point for circles. However, the corresponding reasoning summary initially mentions circles and then abruptly switches to rectangles for no apparent reason (also, the reasoning trace does not make clear how the rectangle size or coordinates were chosen). The reasoning trace is shown below[7].

> **Simulating video frames.** I need to create five image frames to simulate a video. The user mentioned "using five image frames" I'm considering using ASCII art or simple diagrams, maybe with embedded SVG in markdown. For the five frames, I'll create simple shapes – likely circles – and move a smaller shape inside a larger one without touching the boundary initially, then gradually expand it to touch the boundary in the last frame. I'll ensure consistent labeling for clarity.
> **Creating SVG frames for movement.** I'm working on showing the transition where the rectangle x moves until it touches the boundary of rectangle y. This will be done over five SVG frames, each sized 240x160. Rectangle y is set at (20,20) with a size of 200x120, while rectangle x starts at (80,60) with a size of 80x40 and shifts to the right through several frames. In the final frame, x will be at (140,60), ensuring its right edge touches y's right edge at 220. I'll add labels and captions for clarity.

GPT Image 1.5 scored only 1/22 (for PO to TPPi) when generating PNGs and got the starting relation correct in only 11/22 cases. All the diagrams used circles or ellipses. 11/22 diagrams were black and white or gray and white, none were orange and blue. N.B. although we ask for five frames, image generation models such as *gpt-image* can only produce a single raster image depicting the five frames.

[7] OpenAI reasoning traces are usually in Markdown format and so this example has been lightly edited for LaTeX, converting bold text and adjusting layout.

5 Discussion, Conclusions and Future Work

5.1 Contributions

We provide a benchmark that tests diagram recognition (with colour and shape variations), diagram generation, video generation and compositional reasoning using RCC-8. We use it to evaluate multiple contemporary VLMs, LLMs and image/video generation models, and suggest that it could be used to track future progress in multi modal foundation models.

We show that SotA models are able to recognise SVG (vector) diagrams more readily than PNG (raster) diagrams, albeit with problems recognising precise tangents. Diagrams with squares or circles are more readily recognisable than those with triangles or blobs. PNGs are more readily recognised in black and white, or orange and blue rather than green and red or shades of red – this appears to correlate with aspects of the human vision system.

SotA models are able to answer composition questions remarkably well, but there is little evidence that they benefit from explicit prompting for diagrammatic reasoning. For the strongest models there is clearly less scope for improvement but for the weaker models it could have been opportunity.

Although SotA image and video generation models are widely used to, for example, produce social media content, they struggle to provide accurate RCC-8 relation or conceptual neighbourhood illustrations except when SVG is the output format.

5.2 Relation Diagram Recognition

The two best performing VLMs (*GPT-5.2* and *Grok 4*) performed better with SVG recognition than PNG. The models most often confuse tangential relations (TPP, TPPi) with their non-tangential counterparts (NTPP, NTPPi), and sometimes confuse EC with DC. These confusions point to a core difficulty in reliably detecting *boundary contact* – a qualitatively important distinction in RCC-8 that can be difficult to represent precisely in diagrams because of discretisation, anti-aliasing, and stroke width effect. It is possible that increasing the width of the region boundaries (i.e. thicker lines) might help here and remains an area for future work. Although not the most challenging relation to recognise, EQ diagrams are perhaps confusing (the regions are coincident) without the context of seeing other relations such as DC or PO. With SVG recognition, the poor performance with tangents also suggests that the models are not really "visualising" the diagram – they are just looking at the sequence of tokens in the vector representation.

It is perhaps not surprising that relations involving blobs are harder to recognise than squares or circles: blobs combine curved boundaries with irregularity and concavity, which may increase perceptual ambiguity and obscure tangents, as well as possibly requiring more complex calculations in the case of SVG.

Colour seems to affect PNG recognition in a way that roughly mirrors human discriminability: when foreground/background separation is reduced (e.g. adjacent reds), performance collapses. In SVG, where colour is an attribute rather than a noisy percept, performance is comparatively invariant. Taken together, these results suggest that current VLM diagram recognition is still strongly driven by low-level perceptual cues (edge contrast, boundary salience, separability) rather than robust extraction of a representation that is explicitly topological. Others have also reported weakness in VLMs, for example Huynh et al. [18] created a novel benchmark (SalBench) which "consists of images that highlight rare, unusual, or unexpected elements within scenes, and naturally draw human attention" but which VLMs fail to identify.

There is a well known red / green colour vision deficiency in many humans; we hypothesise that because of this many diagrams on the web, and thus potentially in the training data, avoid red / green distinctions, which would presumably result in less training data for VLMs, and thus perhaps might account for the reduced performance we see here, but whether this is really the case cannot be easily ascertained.

Although requiring some interpretation, RCC-8 diagram recognition is largely a recall task and we can reasonably conclude that RCC-8 relation diagrams feature in the training data. We believe that training data generally uses circles rather than squares, triangles or any kind of blob, though there are occasional exceptions: e.g. squares [28]. Thus we can infer some generalisation by the models, or, a good ability to apply the relation descriptions at the start of the prompt to the recognition task.

5.3 Relation Diagram Generation

Like image relation diagram recognition, relation diagram generation is more accurate with SVG than PNG. There are strikingly different styles of image generation, even within one model (Figure 6). Possibly more specific prompting might reduce this variance, but it is still worth noting as seeming rather unhuman-like. Another example of inconsistency surfaces in relation naming: although we defined the inverse relations with a lower case i (TPPi, NTPPi) we sometimes see TPPI and NTPPI (which did appear in some earlier papers on RCC-8, e.g. [13]). In fact, we also see fabricated names like DPP "disconnected proper part" – which is not only not an RCC-8 relation, but also probably a *ridiculous* answer in the sense of Leyton-Brown and Shoham [21] – how can something be both a part and be disconnected?

Image generation models can often create plausible spatial arrangements of shapes, but plausibility is not enough for RCC-8, where correctness depends on crisp boundary conditions. Confusion between TPP and NTPP is especially informative: to depict TPP, the model must place x and y in a position where they touch but do not partially overlap. To depict NTPP, the model must place x strictly inside y with a non-zero margin, avoiding accidental contact under stroke thickness. LLMs are used widely for software engineering and SVGs are effectively a kind of program consisting of vector graphics instructions. It may therefore be easier to specify containment and intersection constraints with explicit coordinates in SVG than to rely on a diffusion-style generator to create topologically correct renditions.

5.4 Compositional Reasoning

None of the evaluated models produce correct results across all CT experiments or even all experimental repeats, however, *GPT-5.2* achieved perfect accuracy (64/64) in one repeat of the CT baseline experiment (with default reasoning effort) and one repeat each of the diagrammatic reasoning and VoT experiments, when reasoning-effort was set to high. The scores are more accurate than those reported in [12] where the best performing models on a similar experiment were *Claude 3.5 Sonnet* (65%) and *Gemini 1.5 Pro* (51%), and these were measured using the more lenient Jaccard Index rather than the strict accuracy we report here. The performance improvement we observe could be indicative of the rapid progress, with new more powerful models, that has been made since 2024 when the paper was published, but may also be influenced by the possible ingestion of [12] as training data.

RCC-8 composition questions can be solved by lookup in the composition table, but vendors do not publish training data and so we cannot be sure whether the models were trained with the RCC-8 CT. If models are simply using recall, one might expect the answers

to be consistently correct, but they are not. Analysis of reasoning traces suggests that some reasoning with topological constraints is being undertaken, but our experiments cannot distinguish these mechanisms conclusively.

If a human was asked to solve CT questions, the subject may well find a reference that includes the full CT table with all relations listed. However there is little evidence of the literature being consulted and we suggest that few humans would have memorised the RCC-8 CT. We searched the *Deepseek V3.2* reasoning traces from the baseline CT experiment for names of authors who have written papers on composition reasoning in RCC-8, and found only infrequent references: Cohn (3/192), Randell (1/92) and Cui (1/92); e.g. "According to standard RCC-8 composition table (see e.g., "Composition Table for the Region Connection Calculus" by Randell, Cui, Cohn)". However the text "known composition table" occurs 37/192 times in the *Deepseek V3.2* reasoning traces for the baseline CT experiment.

Deepseek V3.2 reasoning traces are usually long (median 2492 words for the RCC-8 CT baseline experiment), and we found them hard to interpret. Here is a short example from the simplest composition, EQ(x,y) and EQ(y,z). Note the amount of checking and apparent uncertainty, having first established the (correct) answer:

> We are given: If EQ(x,y) and EQ(y,z) what are all the possible relations that can hold between x and z?
> Recall that EQ is the equality relation. So EQ(x,y) means x = y. Similarly, EQ(y,z) means y = z. By transitivity of equality, we have x = z. Therefore, the only possible relation between x and z is EQ.
> But let's double-check: Could there be any other relation? Since x and y are the same region, and y and z are the same region, then x and z are the same region. So only EQ(x,z) holds.
> Thus, the answer is EQ.
> But the instruction says: "List all possible relations as a comma separated list, with no extraneous text, choosing only from the relations DC, EQ, PO, EC, TPP, TPPi, NTPP, and NTPPi." So we should output exactly "EQ".
> Make sure: Is there any possibility that x and z could be related in another way? For example, if regions are not necessarily exactly the same? But EQ means they are coincident, i.e., identical. So if x = y and y = z, then x = z. So only EQ.
> Thus, the answer is EQ

Contrary to claims made of VoT [30], asking for visualisation or diagram drawing does not reliably improve our results when the model is already strong. We suggest that, since 2024, models have been explicitly trained for reasoning, with "thinking step by step" now being part of alignment. For weaker models, both VoT and explicit diagram instructions produce modest improvements. This indicates that "visualise" prompts can be useful in encouraging models to reflect and reduce some errors. However, the error rate remains too high for tasks where correctness is critical. Moreover, improvements from diagram prompting do not, by themselves, demonstrate true diagrammatic reasoning; they may simply reflect a general benefit of prompt engineering, rather than genuine exploitation of external pictorial representations. It is possible that using different prompting strategies or using k-shot reasoning might genuinely elicit diagrammatic reasoning.

5.5 Conceptual Neighbourhood Illustration

The conceptual neighbourhood (CN) experiments not only require the depiction of two relations correctly, but also a continuous transformation between them that respects the CN transition and preserves identity (labels x and y) over time. This requires temporal coherence, consistent object tracking, and faithful maintenance of qualitative constraints throughout the motion. The poor performance of *Sora 2* and the frequent failure even to depict the starting relation in the first frame, suggests that current video generation models prioritise cinematic plausibility over qualitative spatial constraint satisfaction; we hypothesise that this has not been part of the training regime. While it is remarkable that an AI model can generate videos showing the transformation of shapes at all, these are of little use if they are incorrect. By contrast, when asked for five "frames" as SVG, *GPT-5.2* produced plausible transitions for all CN transitions tested.

5.6 Limitations and Future Work

Many model vendors now publish model-specific prompting guides, for example, OpenAI[8] and xAI[9] and it is possible that targeted prompt engineering might improve the results published here. But to maintain "a level playing field", we elected to use the same prompts throughout.

Although we used three experimental repeats for our colour diagram recognition experiments, VLMs are highly stochastic, and a much larger study is required to further explore whether VLMs really reflect human colour vision deficiency. For the experiments which involved manual evaluation, we only performed a single repeat (n=1), but increasing n would clearly be desirable, and would be straightforward if a way could be found to automate answer checking. Further experiments with different colours and colour combinations might also be interesting.

While we tested diagram recognition using four shapes, it would be interesting to extend the experiment with combinations of shapes and to see whether line width, resolution, aspect ratio, brightness, contrast, blurring or noise affect diagram recognition.

We experimented with PNG and SVG but there there are many other graphics formats. VLM and image generation models are limited in the formats they support (e.g. OpenAI VLMs[10] only support PNG, JPEG, WebP, and non-animated GIF; *gpt-image-1*[11] only PNG, JPEG or WebP). However it would be interesting to try LLMs with textual formats such as Netpbm, PostScript, Graphviz DOT, Mermaid, TikZ, and XAML.

For the recognition and generation experiments we have only experimented with two spatial entities at a time; further experimentation which would challenge models to recognise or generate multiple regions and relations in a single image would be worth conducting.

Although RCC-8 is perhaps the most well known qualitative spatial calculus with relations that encompass connectedness, parthood and tangents, it would be interesting to explore other calculi, including RCC-5 (a simpler calculus that does not consider boundary tangents) and RCC-22 (a more more complex calculus involving concavity and convexity of regions), as well as other qualitative calculi [14].

[8] `https://platform.openai.com/docs/guides/prompt-engineering`, retrieved February 2026
[9] `https://docs.x.ai/docs/guides/grok-code-prompt-engineering`, retrieved February 2026
[10] `https://developers.openai.com/api/docs/guides/images-vision`, retrieved February 2026
[11] `https://developers.openai.com/api/reference/resources/images/methods/generate`, retrieved February 2026

We have noted that *Z Image Turbo* performed poorly in our video generation experiment, but it has been used elsewhere to create photo-realistic images and so the questions arises, could *Z Image Turbo* and other small, open weights video generation models be fine-tuned for QSR illustration?

6 Data Availability

Burnell et al. [7] call for the routine release of LLM evaluation results and more granular reporting practices to improve transparency and robustness in AI system evaluations. Accordingly, we publish all our experimental data including all prompts and API request/response pairs for all experiments[12].

References

1 Dave Barker-Plummer and Sidney C Bailin. The role of diagrams in mathematical proofs. *Machine Graphics and Vision*, 6(1):25–56, 1997.

2 Averi Bates, Ryan Vavricka, Shane Carleton, Ruosi Shao, and Chongle Pan. Unified modeling language code generation from diagram images using multimodal large language models. *Machine Learning with Applications*, 20:100660, 2025. doi:10.1016/j.mlwa.2025.100660.

3 Malcolm I Bauer and Philip N Johnson-Laird. How diagrams can improve reasoning. *Psychological science*, 4(6):372–378, 1993.

4 Jennifer Birch. Worldwide prevalence of red-green color deficiency. *Journal of the Optical Society of America A*, 29(3):313–320, 2012.

5 Robert Blackwell. RCC-8 as a Benchmark for Diagrammatic Reasoning in Multimodal Foundation Models. Dataset, version 1.0., swhId: swh:1:dir:d771bd22f1b7bc35e03e0659eb7f973ccc8fbafb (visited on 2026-08-28). URL: https://github.com/RobBlackwell/cosit2026, doi:10.4230/artifacts.27834.

6 Nasim Borazjanizadeh, Roei Herzig, Eduard Oks, Trevor Darrell, Rogerio Feris, and Leonid Karlinsky. Visualizing Thought: Conceptual Diagrams Enable Robust Planning in LMMs. *arXiv preprint arXiv:2503.11790*, 2025. doi:10.48550/arXiv.2503.11790.

7 Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

8 Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv:2511.22699*, 2025.

9 B. Chandrasekaran, Janice Glasgow, and N. Hari Narayanan, editors. *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. AAAI Press / MIT Press, Cambridge, MA, 1995.

10 An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models, 2024. doi:10.48550/arXiv.2406.01584.

11 Anthony G Cohn, Brandon Bennett, John Gooday, and Nicholas Mark Gotts. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica*, 1(3):275–316, 1997. doi:10.1023/A:1009712514511.

12 Anthony G Cohn and Robert E Blackwell. Can Large Language Models Reason about the Region Connection Calculus? *arXiv preprint arXiv:2411.19589*, 2024. doi:10.48550/arXiv.2411.19589.

[12] https://github.com/RobBlackwell/cosit2026; last acccessed April 23 2026

13 Anthony G Cohn and Nicholas M Gotts. Spatial regions with undetermined boundaries. In *Proc. 2nd ACM Workshop on Advances in Geographic Information Systems*, pages 52–59. Springer, 2011.

14 Anthony G Cohn and Jochen Renz. *Foundations of Artificial Intelligence*, volume 3, chapter Qualitative spatial representation and reasoning, pages 551–596. Elsevier, 2008.

15 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. doi:10.1007/978-3-031-73337-6_9.

16 Eirinaios-Odysseas Gardelakos, Vasileios Kyriakopoulos, Despina-Athanasia Pantazi, Orestis-Minas Kapopoulos, Maria Tsourma, and Manolis Koubarakis. Can Large Reasoning Models Reason about Spatial Relations? In *Proc. 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '25, pages 81–91. ACM, 2025. doi:10.1145/3764912.3770841.

17 Joy Hsu, Gabriel Poesia, Jiajun Wu, and Noah Goodman. Can visual scratchpads with diagrammatic abstractions augment LLM reasoning? In *PMLR 239:21-28*, pages 21–28, 2023. URL: https://proceedings.mlr.press/v239/hsu23a.html.

18 Ngoc Dung Huynh, Phuc H Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. Vision-language models can't see the obvious. In *Proceedings ICCV*, pages 24159–24169, 2025.

19 Mateja Jamnik, Alan Bundy, and Ian Green. Automation of diagrammatic reasoning. *: AAAI Technical Report FS-97-03.*, 1997.

20 Jill H. Larkin and Herbert A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987. doi:10.1016/S0364-0213(87)80026-5.

21 Kevin Leyton-Brown and Yoav Shoham. Understanding understanding: A pragmatic framework motivated by large language models. *arXiv preprint arXiv:2406.10937*, 2024. doi:10.48550/arXiv.2406.10937.

22 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

23 Masataka Okabe and Kei Ito. Color Universal Design (CUD): How to make figures and presentations that are friendly to colorblind people. https://jfly.uni-koeln.de/color/#pallet, 2008.

24 Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. URL: http://proceedings.mlr.press/v139/ramesh21a.html.

25 David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proc. (KR)*, pages 165–176, 1992.

26 Aditi Singh. A survey of AI text-to-image and AI text-to-video generators. In *4th Int. Conf. on AI, Robotics and Control (AIRC)*, pages 32–36. IEEE, 2023.

27 Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. FlowVQA: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1330–1350, 2024. doi:10.18653/V1/2024.FINDINGS-ACL.78.

28 Muralikrishna Sridhar, Anthony G Cohn, and David C Hogg. From video to RCC8: exploiting a distance based semantics to stabilise the interpretation of mereotopological relations. In *International Conference on Spatial Information Theory*, pages 110–125. Springer, 2011. doi:10.1007/978-3-642-23196-4_7.

29 Kristian Tylén, Riccardo Fusaroli, Johanne Stege Bjørndahl, Joanna Raczaszek-Leonardi, Svend Østergaard, and Frederik Stjernfelt. Diagrammatic reasoning: Abstraction, interaction, and insight. *Pragmatics & Cognition*, 22(2):264–283, 2014.

30 Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of LLMs: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024.

31 Shaolin Xie, Shang-Ling Hsu, Qihan Zhang, Yiming Gao, Cyrus Shahabi, and Ibrahim Sabek. Evaluating Intrinsic Geospatial Topological Reasoning in LLMs. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, pages 43–48, 2025. doi:10.1145/3764915.3770722.

32 Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10):101370, 2025. doi:10.1016/j.patter.2025.101370.

33 Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. DiagrammerGPT: Generating open-domain, open-platform diagrams via LLM planning. In *First Conference on Language Modeling*, 2024. URL: https://openreview.net/forum?id=NV8yRJRET1.

34 Enming Zhang, Ruobing Yao, Huanyong Liu, Junhui Yu, and Jiale Wang. First multi-dimensional evaluation of flowchart comprehension for multimodal large language models. *arXiv preprint arXiv:2406.10057*, 2024. doi:10.48550/arXiv.2406.10057.

35 Yifan Zhang, Yang Yuan, and Andrew Chi-Chih Yao. On the diagram of thought. *arXiv preprint arXiv:2409.10038*, 2024. doi:10.48550/arXiv.2409.10038.

An Extended Qualitative Model for Reasoning About 3D Objects Using Depth and Different Perspectives

Martí Ejarque
Computing Science Department, Umeå Universitet, Sweden

Zoe Falomir
Computing Science Department, Umeå Universitet, Sweden

Abstract

Reconstructing a three-dimensional structure from partial two-dimensional views requires spatial reasoning. This paper presents the EQ3D, a dimension-agnostic extension of the Qualitative 3D (Q3D) model that generalises qualitative volumetric reasoning to arbitrary $H \times W \times L$ configurations. EQ3D preserves symbolic depth representations while incorporating adjacent-view propagation rules that increase cross-view inference opportunities. Implemented in SWI-Prolog as a fully declarative pipeline, EQ3D infers hidden views from three orthographic `Front`–`Right`–`Up` views, producing correct inferred cells for the Q3D-Dataset which includes 24 objects of varying size and shape, while ambiguous cells remain explicitly undetermined. Coverage of hidden structure ranges from 26.85% to 82.67%, demonstrating that while inference remains logically sound, the proportion of determinable volume is inherently configuration-dependent. EQ3D provides an interpretable symbolic framework for qualitative 3D reasoning that scales across resolutions without size-specific adaptation.

2012 ACM Subject Classification Theory of computation → Automated reasoning; Theory of computation → Programming logic; Computing methodologies → Knowledge representation and reasoning; Computing methodologies → Spatial and physical reasoning

Keywords and phrases qualitative spatial reasoning, spatial cognition, first order logics, three-dimensional information, perspectives, object perception, SWI-Prolog

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.5

Supplementary Material
Software (Source Code): `https://github.com/Ejarque24/Extended-Q3D` [8]

Funding We acknowledge the funding by the Wallenberg AI, Autonomous Systems and Software Program (WASP) awarded by the Knut and Alice Wallenberg Foundation, Sweden.

1 Introduction

Qualitative Spatial and Temporal Reasoning (QSTR) models properties of space (e.g. topology, orientation, geometry) and their evolution through time. It produces representations and inferences over spatial and temporal configurations at a symbolic level [25, 2, 15]. Rather than relying on precise numerical coordinates, qualitative models capture structural properties such as direction, topology, adjacency, or relative depth [32]. This abstraction allows reasoning under incomplete or imprecise information, while remaining close to human understanding.

The qualitative description of three-dimensional objects and the inference of their spatial structure from partial views is a spatial problem. In many contexts, three-dimensional objects are described through a limited set of canonical projections [30]. Technical drawings and some educational spatial tasks present three orthogonal views (typically front, lateral and top) to represent an object [34] and ask to infer the occluded views. The inferential strength of some descriptions varies significantly: some sets of projections might constrain the hidden

© Martí Ejarque and Zoe Falomir;
licensed under Creative Commons License CC-BY 4.0

structure of the object, while others can leave substantial ambiguity [23]. Understanding what makes a qualitative 3D description informative, and how much of the hidden structure can be reconstructed from it, is our research question.

The Qualitative 3D depth model (Q3D [10, 37]) addressed this challenge by introducing a symbolic representation of depth levels across three canonical views. In Q3D, each projection is described as a matrix of qualitative depth symbols, and logical consistency constraints govern the relations between opposing and adjacent views. The model demonstrated that hidden views can be partially reconstructed through symbolic reasoning alone, without recourse to metric geometry. In its original formulation, Q3D was defined over a fixed $3 \times 3 \times 3$ qualitative partition. This discretisation provides a conceptually transparent substrate for expressing depth relations and inter-view constraints. However, if qualitative depth reasoning is to apply to objects of varying size, the underlying constraints must be formulated independently of a specific dimension instantiation. Achieving such independence requires a parametrised reformulation in which depth semantics and cross-view consistency are invariant under arbitrary height (H), width (W), and length (L) dimensions, $H \times W \times L$.

Thus, this paper presents the Extended Qualitative (EQ3D), a dimension-agnostic reformulation of the Q3D framework, which generalises the qualitative volumetric description from a fixed $3 \times 3 \times 3$ partition to arbitrary $H \times W \times L$ dimensions, while preserving the ordinal interpretation of depth symbols and the logical structure of inter-view constraints. The extension is not achieved by merely scaling the original construction, but by redefining adjacency relations and boundary correspondences in an index-independent manner. As a result, the qualitative semantics of depth and consistency become invariant under changes in resolution and aspect ratio. By establishing this dimensional independence, EQ3D provides a unified qualitative framework capable of describing objects of arbitrary size within a single formal system. In addition, the generalised formulation makes the propagation of constraints across views explicit, enabling a systematic analysis of how structural configurations influence the reconstruction of hidden perspectives.

The rest of this paper is organised as follows. Section 2 describes related work. Section 3 outlines the Q3D model. Section 4 describes the Extended Q3D model and its inference mechanisms. Section 5 report the experimental results and coverage analysis. Section 6 discusses the results and structural factors influencing inferential reach. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Related Work

Qualitative models [14] have been developed to solve cognitive spatial reasoning tests, such as object sketch recognition tests [27], oddity tasks [28], Raven's Progressive Matrices [29], paper folding reasoning tests [9, 13], qualitative shape composition [12, 31] and object rotation tests [11]. As far as we are concerned, there is no qualitative model developed to reason about the volumetric description of a generalized 3D object of $H \times W \times L$. In the field of Qualitative Spatial and Temporal Reasoning (QSTR), the most relevant calculi was reported by Dylla et al. [7], and there are no calculi that reasons about volumes, that we are aware of.

In the field of computer graphics and geometrical modeling, reconstructing three-dimensional structures/objects from multiple orthographic views has been studied extensively [16, 19]. Early symbolic and geometric approaches assume precise correspondences between views and rely on deterministic constraints to recover a unique 3D model [35]. Line-drawing interpretation and projection-consistency methods, for instance, compute spatial structure through exact geometric reasoning over edges, views, and junctions [21, 26, 38]. While effective under idealised conditions, such approaches presuppose metric precision and are not designed for qualitative abstraction or symbolic depth representations.

Some non-probabilistic "soft" approaches have also explored rule-based treatments of ambiguity while reconstructing 3D objects. Gorgani et al. [17] proposed a fuzzy surface analysis framework that, like EQ3D, relies on expert-defined rules rather than purely numeric optimisation. However, whereas their model represents uncertainty through graded membership within a fuzzy setting, EQ3D adopts a fully qualitative symbolic semantics in which ambiguity arises from logical underdetermination.

Recent work has reframed orthographic reconstruction as a data-driven inference problem. SPARE3D [18] introduces a benchmark for spatial reasoning over three-view drawings, including tasks such as view consistency checking, pose recognition, and shape generation. While these tasks are closely related to EQ3D in that it also starts from `Front-Right-Up` views, SPARE3D relies on learned statistical patterns rather than explicit symbolic constraints. Notably, standard deep neural models perform close to random guessing on several reasoning tasks [18], highlighting the difficulty of capturing structural spatial constraints implicitly. Because SPARE3D does not provide qualitative depth encodings, it cannot be directly used within the EQ3D reasoning framework.

PlankAssembly [20] employs transformer-based architectures to predict "shape programs" for block-like CAD objects from orthographic views, similarly to EQ3D. By learning structural priors and encoding constraints implicitly in the decoding process, the system achieves robustness to noise and incomplete inputs. However, unlike EQ3D's explicit symbolic rules and declarative queryable inference mechanisms, its consistency conditions and inference rules remain embedded in model parameters and not directly explainable or inspectable.

Hybrid neuro-symbolic approaches have also been explored in block-based reasoning domains [24, 4]. Krishnaswamy, Friedman, and Pustejovsky [24] combined qualitative spatial relations with learning and heuristic search to acquire structural concepts from sparse and noisy examples. However, there is not a neuro-symbolic model for constructing 3D objects from 2D views, as far as we are concerned.

3 Preliminaries: the Qualitative 3D depth model

When reasoning about real objects in space, humans naturally think in terms of three-dimensional volumes and multiple perspectives. The Q3D model [10] formalises this intuition by encoding orthographic projections through symbolic depth values rather than numeric measurements. The Q3D depth Descriptor [10, 37] of an object is defined by three views:

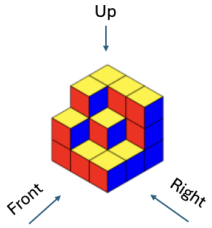
$$\begin{aligned} &view(F, Object, H, W, Q3Dmatrix(H, W)),\\ &view(R, Object, H, L, Q3Dmatrix(H, L)),\\ &view(U, Object, L, W, Q3Dmatrix(L, W)). \end{aligned}$$

where F, R, and U denote the `Front`, `Right`, and `Up` perspectives, while H, W, and L represent the height (H), width (W), and length (L) of the object's bounding box. Each matrix entry belongs to the symbolic set:

$$Q3Dmatrix(D_1, D_2) \subseteq \{a, b, c, d, \ldots, *\}.$$

where D_1 and D_2 are the dimensions of the matrix (i.e. pairs of 3D dimensions: H, W and H, L and L, W) and the elements of each matrix describe ordinal depth layers composed of equally sized volumetric units: a denotes the surface of the object, b the first level of depth (one unit removed), c the second level of depth (two units of volume removed), d the third level of depth (three units of volume removed) and so on. The symbol $*$ indicates that all volumes along a depth column have been removed, so there is a hole. These symbols encode relative ordering rather than metric distance.

Figure 1 presents an object described by the Q3D where volumes of a specific size produced a $3 \times 3 \times 3$ grid with volumetric features for the `Front`, `Right`, and `Up` sides. Note that another volumetric size could have been used producing other dimensions for the object descriptor (e.g. $6 \times 6 \times 6$).



Q3D descriptor for this 3x3x3 object:

`Front`	`Right`	`Up`
$[b, c, c]$	$[*, c, a]$	$[a, a, a]$
$[a, b, c]$	$[c, b, a]$	$[a, b, c]$
$[a, a, a]$	$[a, a, a]$	$[b, c, c]$

Logically as Prolog facts:

$view(front, obj, 3, 3, [[b, c, c], [a, b, c], [a, a, a]]).$
$view(right, obj, 3, 3, [[*, c, a], [c, b, a], [a, a, a]]).$
$view(up, obj, 3, 3, [[a, a, a], [a, b, c], [b, c, c]]).$

Figure 1 Example of (a) an object observed from `Front`, `Right`, and `Up` sides; (b) an intuitive representation of the Q3D descriptor; (c) the logic facts produced by the Q3D descriptor for this specific object.

Consistency Checking across Views

Beyond representation, Q3D imposes geometric compatibility constraints between views. A Q3D description is *consistent* [10] if all pairwise projections can be jointly realised as orthographic views of a single three-dimensional object:

$$\begin{aligned} consistent_Q3D(F, R, U) \leftrightarrow\ & consistent_perspective(F, R) \wedge \\ & consistent_perspective(F, U) \wedge \\ & consistent_perspective(R, U). \end{aligned}$$

Intuitively, two views are consistent if there exists (at least) a 3D arrangement of volumetric blocks whose projections simultaneously satisfy the qualitative depth relations encoded in both views. Thus, `Front`–`Right` consistency requires that the volume values in last column of the `Front` view are consistent with the depth columns of the `Right` view, since they meet in an object edge and therefore these volumes are the same seen from both views. `Front`–`Right` consistency implies `Right`–`Front` consistency, which is defined as:

$$\begin{aligned} F_{i,3}(a) &\leftrightarrow R_{i,1}(a) \\ F_{i,3}(b) &\leftrightarrow R_{i,2}(a) \wedge (R_{i,1}(b) \vee R_{i,1}(c) \vee R_{i,1}(*)) \\ F_{i,3}(c) &\leftrightarrow R_{i,3}(a) \wedge (R_{i,1}(b) \vee R_{i,1}(c) \vee R_{i,1}(*)) \wedge (R_{i,2}(b) \vee R_{i,2}(c) \vee R_{i,2}(*)) \\ F_{i,3}(*) &\leftrightarrow \bigwedge_{k=1}^{3} \neg R_{i,k}(a). \end{aligned}$$

where $\forall k, i \in \{1, 2, 3\}$. Similarly, `Front`–`Up` consistency and `Right`–`Up` consistency are defined.

Inference of Depth Values in Occluded Object Sides

The Q3D model also provided deterministic rules for inferring depth values in the occluded sides of an object: `Back`, `Left`, and `Down`. For that, these rules exploit space continuity and depth interrelations. For example, `Front` and `Back` are opposite sites, if the Q3D include an item indicating the last level of depth in a view (e.g., level c in `Front` in an object of $3 \times 3 \times 3$ dimension), that involves the existence of the first level of depth (a) in the opposite view,

that is `Back`. The same interrelations appear in opposite `Right` and `Left` sides and in `Up` and `Down`. Regarding space continuity, the case of holes involves that symbols ($*$) propagate from-to opposite sides of an object.

In summary, the Q3D depth model introduced: (i) a symbolic volumetric representation, (ii) cross-view consistency constraints, and (iii) inference mechanisms based on space continuity (e.g. hole continuity) and interrelations (e.g. depth relativity). It successfully showed that spatial structure and depth relations can be inferred from complementing perspectives. The Q3D reasoning model was implemented and tested in a SWI-Prolog application using 3×3×3 objects [10] and then it was implemented in a videogame [37] with the goal of training players' spatial reasoning skills.

In this paper, we extend the Q3D model to any dimension and we also include a new inference rule for discovering volumes in occluded sides which is based on conceptual neighbourhood between sides.

4 The Extended Qualitative 3D Depth Model (EQ3D)

The Extended Qualitative 3D depth model (EQ3D) builds directly upon the foundations established by the original Q3D model [10] by providing descriptors for objects of any dimension $H \times W \times L$, preserving the qualitative initial representation while extending its expressive power to a broader range of object geometries. This generalization enables finer volumetric partitions and supports shapes that are elongated, asymmetric, or otherwise more detailed than those captured by the original configuration.

Specifically, EQ3D generalizes to arbitrary object size ($H \times W \times L$) while preserving the underlying qualitative depth semantics, it propagates depth information using adjacency-based cross-view constraints across neighbouring views.

Moreover, it exploits adjacency relations between views to further extend the inference capabilities of Q3D by using the consistency checking rules as inference rules discovering volumes existing in the occluded neighbouring sides (e.g. for inferring new depth values for `Back`, we can check the views `Right` vs. `Back`, `Left` vs. `Back`, `Down` vs. `Back` and `Up` vs. `Back`, specifically at the edges of the object, where the volumes seen are the same.) These new rules exploit local depth continuity across shared boundaries, allowing information to propagate between neighbouring views. In doing so, EQ3D complements and enriches the original inference framework, extending qualitative reasoning towards more complex and realistic 3D interpretations. These generalizations provide a unified qualitative framework that supports both consistency checking and the reconstruction of hidden views from partial input.

Notation and Preliminaries. EQ3D retains the original formal foundations of Q3D, where three canonical perspectives (`Front` (F), `Right` (R), and `Up` (U)) encode qualitative depth using symbolic letters. Each perspective is expressed as a matrix whose entries belong to: $\Sigma = \Sigma_d \cup \{*\}$, $\Sigma_d = \{a, b, c, d, \dots\}$ where a denotes the nearest visible surface, deeper levels follow the alphabetical order, and $*$ represents enforced absence of volume.

The qualitative structure is now generalized to an arbitrary grid of dimensions (H, W, L), corresponding respectively to the vertical, horizontal, and depth axes. Each axis is discretized into uniform volumetric units, without fixing the number a priori, allowing the model to adjust to the complexity of the described object.

Formally, a perspective P is a matrix of depth symbols: $P = [p_{i,j}]$, $p_{i,j} \in \Sigma$, with dimensions determined by how the view aligns with the (H, W, L) grid.

▶ **Definition 1.** *Let the qualitative reference system of EQ3D be:*

$$EQ3D^{(H,W,L)} = \left\{P_v \mid v \in V,\ P_v \subseteq \Sigma^{n_v \times m_v},\ (n_v, m_v) \in \{(H,W),(H,L),(L,W)\}\right\},$$

representing all canonical perspectives with their geometrically determined sizes.

The depth alphabet Σ_d encodes only ordinal information (near/far ordering), independent of the chosen grid resolution. This separation ensures that qualitative reasoning is scale-invariant across different object sizes.

Generalization to Arbitrary Dimensions $H \times W \times L$. The EQ3D generalizes dimensions by expressing all adjacency relations between views in an index-agnostic manner: the rules refer only to the boundary indices that two views share, and therefore remain valid independently of the absolute size or dimension.

In the EQ3D qualitative partition, each canonical view corresponds to a projection onto one of the coordinate planes. Adjacent views meet along a one-dimensional boundary, and the structure of this boundary determines how their symbolic descriptions must agree. For example, the rightmost column of the `Front` view coincides with the leftmost column of the `Right` view; similarly, the top row of the `Front` view aligns with the bottom edge of the `Up` view; and the upper row of the `Right` view aligns with the rightmost edge of the `Up` view, subject to the natural mirroring imposed by the viewing direction.

▶ **Definition 2** (Neighbouring boundary alignments)**.** *Let $F \in N_{\text{depths}}^{H\times W}$ and $R \in N_{\text{depths}}^{H\times L}$ denote the* ***Front*** *and* ***Right*** *views. For each row i, the boundary cells $F(i,W)$ and $R(i,1)$ describe the same vertical line of the 3D object. If $F(i,W) = *$, then the entire $R(i,1)$ row must be free of volume units, since there is a hole on the* ***Front*** *view at that location which excludes the possibility of a visible surface from the* ***Right*** *view. Conversely, if $F(i,W) \neq *$, then row i of the* ***Right*** *view must show a surface at exactly the depth index associated with that qualitative depth value. A symmetric constraint maps the leftmost column of R back to the appropriate mirrored column of F, ensuring that both views encode the same shared edge consistently. The remaining view adjacency pairs in the object follow the same principle, thus:*

$$\begin{aligned} &F(i,W) \leftrightarrow R(i,1), && i = 1,\dots,H,\\ &F(1,j) \leftrightarrow U(1,j), && j = 1,\dots,W,\\ &R(1,k) \leftrightarrow U(k,W), && k = 1,\dots,L. \end{aligned}$$

That is, the top row of the `Front` view shares its boundary with the nearest edge of the `Up` view; thus, each symbol $F(1,j)$ constrains the corresponding column of U when read from bottom to top. Likewise, the upper row of the `Right` view meets the rightmost column of the `Up` view; here, the alignment requires horizontal mirroring due to the opposite viewing direction, but the constraint itself is structurally identical to the previous cases. Since the rules quantify only over these boundary indices, and not over fixed-size patterns, then the adjacency constraints automatically generalize to any choice of (H,W,L). Thus, EQ3D preserves the structure of Q3D consistency reasoning while allowing objects of arbitrary resolution and aspect ratio to be described and validated.

Computational implications. Checking pairwise edge consistency across the three adjacent pairs ($F \leftrightarrow R$, $F \leftrightarrow U$, $R \leftrightarrow U$) requires a number of comparisons proportional to the lengths of shared edges. Summed over all rows/columns, this yields time linear in the total number of view cells, $\mathcal{O}(H \cdot W + H \cdot L + L \cdot W)$. When inference propagates information throughout

the interior of the volume (e.g., to reconstruct occluded cells or to run completion on all views), the work scales linearly with the number of volumetric units, $\mathcal{O}(H \cdot W \cdot L)$. Therefore, EQ3D preserves linear scaling with problem size at both the surface (views) and volume levels.

Illustrative example (4×4×3). Consider an object discretized over an EQ3D grid with dimensions $H = 4$ (Height, Cartesian axis y), $W = 4$ (Width, Cartesian axis x), and $L = 3$ (Length, Cartesian axis z). Figure 1 shows the qualitative description of such an object using its three canonical perspectives: `Front` (4×4), `Right` (4×3), and `Up` (3×4).

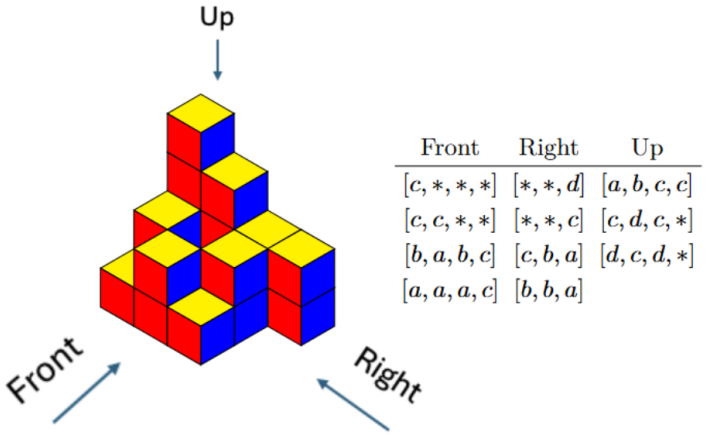


Figure 2 Example of a discretized 3D object showing the `Front`, `Right`, and `Up` views.

Figure 2 shows the EQ3D descriptor for an object of size: 4×4×3. The `Front` view specifies, for each of the 4 rows and 4 columns, the nearest visible depth letter along the front-to-back axis (i.e. Cartesian z axis). The `Right` view provides the same information from the right-hand side: for each of the 4 rows and 3 columns, it encodes the nearest visible depth along the right-to-left axis (i.e. Cartesian x axis). The `Up` view, of size 3×4, corresponds to looking from above (i.e. Cartesian y axis): its 3 rows index depth layers from nearest to farthest, while its 4 columns align with the horizontal positions shared with the Front view.

4.1 EQ3D Consistency Checking

The EQ3D model checks that the `Front`, `Right` and `Up` descriptors are logically coherent between them to produce a robust description of a single 3D object. This is done by enforcing three layers of consistency constraints: well-formed views, object completeness and cross-view consistency along shared edges.

A view is well-formed if all rows have the same length, its declared dimensions (*rows*, *cols*) match the actual matrix dimensions, and every entry belongs to N_{depths}. Views that violate any of these conditions are rejected as invalid descriptions and cannot be used for EQ3D reasoning. Completeness is considered when all size canonical perspectives are available. In practice, our focus is on the three input views F, R, U, while B, L, D is used as ground truth for the later inferred views. However, whenever all six views are provided completeness checking ensures that no view is missing and no extra/non-canonical label is present. This guarantees that all consistency rules below are defined on a fixed, common set of perspectives.

Adjacency-based consistency. The core of EQ3D consistency checking is the set of *adjacency rules* that relate pairs of neighbouring views along their shared edges. Conceptually, these rules state that whenever two views touch in 3D, their boundary cells must encode compatible depth information. EQ3D checks three adjacency pairs, each in both directions: (F, R), (F, U), (R, U) yielding the six constraints FR, RF, FU, UF, RU, and UR. In all rules below, we assume that each depth letter $\ell \in \{a, b, c, \dots\}$ is mapped to a *depth index* $\mathrm{idx}(\ell) \in \mathbb{N}$, starting at 0 for a, 1 for b, and so on.

▶ **Definition 3** (EQ3D Consistency Front → Right)**.** *For each row i, the last column of the Front view (rightmost edge) meets the first column of the Right view (leftmost edge). The rule is:*

(FR-)* $F(i, W) = * \;\Rightarrow$ *no cell* $R(i, \cdot)$ *in that row is* a,
(FR-a) $F(i, W) = \ell \neq * \;\Rightarrow\; R\big(i, \mathrm{idx}(\ell)\big) = a.$

Intuitively, if the last visible point in `Front` is a hole, `Right` must not see a surface a along that row. If instead Front sees depth ℓ, then Right must show a surface at the corresponding depth index.

▶ **Definition 4** (EQ3D Consistency Front ← Right)**.** *Symmetrically, we check the first column of Right against the entire row of Front, but now taking into account horizontal mirroring. Let W be the number of columns of Front, and* $\mathrm{mir}(k) = W - 1 - k$ *the mirrored index. Then let* $\mathrm{mir}(k) = W - 1 - k$ *denote the horizontal mirroring index for a view with W columns. For each row i:*

(RF-)* $R(i, 1) = * \;\Rightarrow$ *no cell* $F(i, \cdot)$ *in that row is* a,
(RF-a) $R(i, 1) = \ell \neq * \;\rightarrow\; F\big(i, \mathrm{mir}(\mathrm{idx}(\ell))\big) = a.$

FR and RF together ensure that Front and Right encode the same edge in a consistent way, up to the natural mirroring of the viewing axes.

▶ **Definition 5** (EQ3D Consistency Front → Up)**.** *The top row of Front shares its boundary with the front edge of the Up view. For each column j, we compare $F(1, j)$ with the column of Up at position j when read from bottom to top:*

(FU-)* $F(1, j) = * \;\Rightarrow$ *no cell in the corresponding Up column is* a,
(FU-a) $F(1, j) = \ell \neq * \;\Rightarrow$ *the Up column at j has a at depth index* $\mathrm{idx}(\ell)$.

Thus, a hole at the top of `Front` discards surfaces directly above it in `Up`, whereas a visible depth letter demands a matching surface at the corresponding vertical dimension Length (Cartesian axis z) in the Up view.

▶ **Definition 6** (EQ3D Consistency Front ← Up)**.** *Conversely, the bottom row of Up is aligned with the full height of Front. For each column j:*

(UF-)* $U(L, j) = * \;\Rightarrow$ *no cell in the Front column j is* a,
(UF-a) $U(L, j) = \ell \neq * \;\rightarrow\; F\big(\mathrm{idx}(\ell) + 1,\, j\big) = a.$

Here, a hole at the bottom of Up excludes any visible surface in the corresponding Front column, while a depth letter fixes the row at which Front must show a surface.

▶ **Definition 7** (EQ3D Consistency Right → Up)**.** *The top row of Right is adjacent to the right edge of Up. For each depth index that is simultaneously valid for both views, the rule is:*

(RU-)* $R(1, k) = * \;\Rightarrow$ *the corresponding mirrored row of Up has no* a,
(RU-a) $R(1, k) = \ell \neq * \;\Rightarrow$ *the corresponding mirrored row of Up contains a at index* $\mathrm{idx}(\ell)$.

Each position on top edge of Right must be compatible with the row of Up that represents the same depth, read with the appropriate horizontal mirroring.

▶ **Definition 8** (EQ3D Consistency Right ← Up)**.** *The rightmost column of Up is matched with a family of depth-indexed columns in Right. For each row i of Up (which corresponds to a depth layer), we enforce:*

$$\text{(UR-*)} \quad U(i, W_U) = * \;\Rightarrow\; \text{no } a \text{ occurs in the corresponding depth column of Right},$$
$$\text{(UR-}a\text{)} \quad U(i, W_U) = \ell \neq * \;\Rightarrow\; \text{the Right column at depth } \mathrm{idx}(\ell) \text{ contains } a \text{ at row } i.$$

An EQ3D object is said to be adjacency-consistent if all rules (Front ↔ Right ∧ Front ↔ Up ∧ Right ↔ Up) are satisfied simultaneously. In logical terms, each rule is a local constraint over pairs of boundary cells, but their conjunctions ensure that the three given views F, R, U could arise from a single coherent 3D configuration. This global consistency is a prerequisite for the following inference stage, which uses, to some extent, the same qualitative set and adjacency relations to complete the hidden views B, L, D.

4.2 EQ3D Reasoning Inferences

Assuming the orthographic views `Front`, `Right` and `Up` are provided in an EQ3D description, we ask: how much of the remaining object can be deduced logically? In particular, can the `Back`, `Left` and `Down` perspectives be partially computed from the given views?

Each view is represented by a matrix of symbols over $\{a, b, \ldots\} \cup \{*\}$, where letters encode depth along a sightline (a nearest to the observer) and `*` denotes a hole (an enforced absence). Unknown or not yet inferred entries are written as "_":

EQ3D reasons about hidden views using three families of qualitative inference rules: opposite-view depth relativity, opposite-view hole continuity, and additional adjacent depth cues that directly constrain `Back`, `Left` and `Down`.

4.2.1 Opposite-view depth relativity

Depth-relativity captures the idea that the deepest visible symbol along one axis guarantees the presence of a surface in the opposite view. EQ3D distinguishes three depth axes:

- deep_{FB}: deepest letter along the `Front`/`Back` axis (i.e. Cartesian z axis).
- deep_{LR}: deepest letter along the `Left`/`Right` axis (i.e. Cartesian x axis).
- deep_{UD}: deepest letter along the `Up`/`Down` axis (i.e. Cartesian y axis).

Intuitively, depth relativity propagates guaranteed occupancy, hole continuity propagates guaranteed absence, and adjacent depth cues exploit shared boundaries to refine the structure of hidden views. For a given object, the maximum number of depth levels are determined based on the dimension of the axis (e.g. if there are four levels, $\mathrm{deep}_{LR} = d$).

▶ **Definition 9** (Inference Depth Relativity Front ↔ Back)**.** *Let F be the Front view with N columns, and B the Back view. For each row i and column j:*

$$F(i, j) = \mathrm{deep}_{LR} \;\leftrightarrow\; B\big(i, \mathrm{mir}(j)\big) = a, \quad \mathrm{mir}(j) = H - 1 - j.$$

That is, whenever Front sees the deepest possible depth at position (i, j), the mirrored column $\mathrm{mir}(j)$ in the Back row i must contain a visible surface a.

▶ **Definition 10** (Inference Depth Relativity Right ↔ Left). *Symmetrically, let R be the Right view with N columns, and L the Left view. For each row i and column j:*

$$R(i,j) = \text{deep}_{LR} \;\leftrightarrow\; L\big(i, \text{mir}(j)\big) = a.$$

Again, a deepest letter on Right at (i,j) enforces a visible surface on Left at the mirrored position of the same row.

▶ **Definition 11** (Inference Depth Relativity Up ↔ Down). *For the vertical axis, let U be the Up view with K rows and M columns, and D the Down view. For each Up cell (i,j):*

$$U(i,j) = \text{deep}_{UD} \;\leftrightarrow\; D\big(\text{mir}_v(i), j\big) = a, \quad \text{mir}_v(i) = L - 1 - i.$$

Here, a deepest letter seen from above at row i, column j implies a surface at the mirrored vertical depth $\text{mir}_v(i)$ in the Down view at the same horizontal position j. Intuitively, if the object is "full" as far as the Up view can see at that sightline, then the opposite Down view must also encounter a surface at the extreme depth.

These depth-relativity rules never introduce holes; they only add certain a cells on the opposite view, reflecting the idea that reaching the last usable depth from one side guarantees occupancy when seen from the other side.

4.2.2 Opposite-view hole continuity

Hole continuity states that a hole observed from one side of the object must also appear as a hole in the directly opposite view, at the mirrored coordinate. This captures the idea that an enforced absence of volume cannot be "filled" when seen from the other side. For each axis, the deepest letter is determined by the maximum number of qualitative depth levels implicit in the corresponding view dimensions. For instance, if four levels are used along the left–right axis, then $\text{deep}_{LR} = d$.

▶ **Definition 12** (Inference Hole Propagation Front ↔ Back). *Let F have H columns. For every cell (i,j):*

$$F(i,j) = * \;\leftrightarrow\; B\big(i, \text{mir}(j)\big) = *.$$

▶ **Definition 13** (Inference Hole Propagation Right ↔ Left). *Similarly, for Right and Left:*

$$R(i,j) = * \;\leftrightarrow\; L\big(i, \text{mir}(j)\big) = *.$$

▶ **Definition 14** (Inference Hole Propagation Up ↔ Down). *For Up and Down, using vertical mirroring:*

$$U(i,j) = * \;\leftrightarrow\; D\big(\text{mir}_v(i), j\big) = *.$$

Opposite view hole continuity and depth-relativity are complementary: the former propagates enforced absences ($*$), while the latter propagates guaranteed surfaces (a) from the deepest observable level. Both act monotonically: they never remove information, only refine unknown cells "_" into either $*$ or a.

4.2.3 Additional adjacent depth cues

Beyond the opposite-view constraints, EQ3D incorporates a set of adjacent depth cues that arise from views sharing not only edges but also orthographic sightlines whose extremal rows or columns encode depth information. These cues exploit the fact that certain boundary positions (such as the last column of the Right view, the top row of the `Up` view, or the leftmost column of `Up`) correspond directly to depth indices in the hidden views. As a result, specific depth letters in these boundary positions can be interpreted as direct constraints on the `Back`, `Left`, or `Down` views.

Concretely, a depth letter appearing in an extremal row or column of a visible view fixes the location of a surface cell in the corresponding hidden view. For instance, the last column of the **Right** view represents the furthest depth positions along the right-to-left axis; thus, if $R(i, L) = \ell \neq *$, the depth index $\text{idx}(\ell)$ specifies the column of the `Back` view in which a surface must appear at row i: $R(i, L) = \ell \neq * \;\rightarrow\; B\big(i, \text{idx}(\ell)\big) = a$.

Analogous forms of depth-index projection apply to the top row of the `Up` view (informing `Back`), the leftmost column of `Up` (informing `Left`), and the bottom rows of `Right` or `Front` (informing `Down`). These cues do not replace the adjacency rules but enrich them by providing direct depth-to-position mappings that become available whenever the relevant boundary cells contain explicit depth letters. Together with the previous inference mechanisms, they contribute to a more complete and deterministic reconstruction of the hidden views B, L, and D.

5 Results

Dataset and Experimental Setup

The EQ3D model was implemented as first-order Horn clauses using SWI-Prolog[1] as the testing platform [36] following a fully declarative inference pipeline. All experiments were executed under identical symbolic parameters, without size-dependent hardcoding, ensuring generality across $H \times W \times L$ configurations.

Then it was evaluated on the Q3D-Dataset, which contains 24 objects ranging from compact cubic configurations ($2 \times 2 \times 2$) to larger rectangular shapes (up to $6 \times 7 \times 6$) which was originally created for the Q3D-videogame [37] and it is a compilation of objects from the literature [1, 3, 34, 10]. Each object was specified through six canonical perspectives (`Front`, `Right`, `Up`, `Back`, `Left`, `Down`), following the EQ3D reference system introduced in Section 4.

For each object, only the three visible views (`Front`, `Right`, `Up`) were provided as input facts. The remaining three hidden views (`Back`, `Left`, `Down`) were inferred automatically by the EQ3D model using the adjacency and depth-consistency rules defined in Section 4.1. Objects were normalised prior to inference to ensure canonical alignment and to avoid ambiguities due to symmetry or rotation. All experiments were executed under identical symbolic parameters, without size-dependent hardcoding, ensuring generality across $H \times W \times L$ configurations. Both the data and the code for this work are available here.[2]

[1] SWI-Prolog: `http://www.swi-prolog.org/`
[2] `https://github.com/Ejarque24/Extended-Q3D.git`

Evaluation Protocol

We evaluated the EQ3D model using two measures: *correctness* and *inferential completeness*. Correctness is assessed by comparing every inferred hidden-view cell against its ground-truth qualitative depth value for each object. Inferential completeness is assessed using a coverage measure which is defined over the hidden views only (`Back`, `Left`, `Down`), and reported alongside compact inference indicators.

Correctness. Across all objects, inferred cells match the ground-truth qualitative descriptors with **100% accuracy**. This follows from the soundness of the EQ3D inference rules: the system derives only information that logically follows from the visible view together with volumetric continuity and depth relativity. No contradictions or inconsistent configurations were produced.

Coverage as Inferential Completeness. To quantify inferential reach, we measure how much of the hidden structure (`Back`, `Left`, `Down`) can be determined from the visible views. We defined coverage as the proportion of hidden-view cells whose values become determined (either a depth symbol *a,b,c,...* or a hole `*`). The hidden-view cells are composed of all the unseen cells in `Back`, `Left`, `Down`, including the boundary cells shared with the shown views:

$$\text{Coverage}(O) = \frac{\texttt{determined hidden cells}(O)}{\texttt{total hidden cells}(O)} \tag{1}$$

Thus, coverage quantifies *inferential completeness*, independently of correctness. Coverage values range from **26.85%** (object 18) to **82.67%** (object 9), with a dataset mean of approximately 60%. This confirms that while inference is uniformly correct, the proportion of reconstructable hidden structure varies substantially across configurations.

To contextualise coverage, we report three compact global indicators that capture inferential behaviour and visible structural properties. Table 1 shows representative objects spanning the coverage spectrum. The reported indicators are defined as follows:

- **Collision Rate (CR):** inferential redundancy, measured as the proportion of inference events that do not expand the set of determined hidden cells (i.e., multiple rule firings targeting already-determined cells). Lower values indicate broader spatial dispersion of inference over distinct target cells.
- **Boundary Share (BSS):** the proportion of inference events triggered by the dedicated inference-boundary rules (Right→Back last-column, Up→Back top-row, Up→Left left-column, Right→Down last-row, Front→Down last-row). High BSS indicates that inference is dominated by boundary-propagation rules rather than interior propagation.
- **INER (Interior Non-Extreme Ratio):** the proportion of visible cells *excluding inference-boundary cells* whose values are neither holes (`*`) nor view-dependent maximum depth symbols. INER provides a compact measure of how much interior visible structure lacks strong extremal cues.

Coverage Spectrum

Objects cluster into three empirical types:

- **High coverage** ($\geq 70\%$; 7 objects): objects 5, 9, 10, 11, 12, 13, and 21, with object 9 reaching 82.67%.
- **Medium coverage** (50–70%; 9 objects): objects 1, 2, 3, 4, 6, 7, 8, 14, 16, 17, and 24.
- **Low coverage** ($< 40\%$; 4 objects): objects 18, 20, 22, and 23.

Table 1 Coverage and selected inference indicators.

Object	Origin	Coverage (%)	Collision Rate (CR)	Boundary Share (BSS)	INER
o1	Q3D [10]	55.56	0.3750	0.6250	0.4286
o2	Q3D [10]	59.26	0.3846	0.5769	0.2857
o3	Q3D [10]	55.56	0.1176	0.8824	0.8571
o4	Q3D [10]	51.85	0.1765	0.8824	0.9286
o5	DST_01 [34]	74.07	0.3548	0.4516	0.0714
o6	DST_02 [34]	55.56	0.2500	0.4500	0.6429
o7	DST_03 [34]	66.67	0.2800	0.3600	0.3571
o8	PAT_5A [3]	66.67	0.1351	0.2162	0.4333
o9	PAT_5B [3]	82.67	0.0746	0.1642	0.1923
o10	PAT_5C [3]	79.17	0.2830	0.3585	0.0333
o11	PAT_5D [3]	72.92	0.1463	0.2927	0.3000
o12	PAT_5Z [3]	75.00	0.5263	0.5263	0.0000
o13	PAT_100 [3]	76.00	0.1618	0.3529	0.3462
o14	PAT_113 [3]	58.33	0.2821	0.3590	0.3333
o15	PAT_222 [3]	48.15	0.1333	0.8667	1.0000
o16	PAT_224 [3]	61.33	0.0800	0.1800	0.4808
o17	PAT_225 [3]	68.75	0.1316	0.4211	0.4000
o18	PAT_230 [3]	26.85	0.0333	0.5333	0.9750
o19	PAT_219 [3]	49.33	0.0750	0.6000	0.7115
o20	PAT_114 [3]	38.27	0.0882	0.7647	0.9074
o21	PAT_200 [3]	79.17	0.0952	0.8095	0.5556
o22	PAT_220 [3]	38.55	0.0857	0.6286	0.8621
o23	PAT_221 [3]	38.89	0.0667	0.8000	0.9167
o24	PAT_689 [3]	52.78	0.0952	0.7143	0.8889

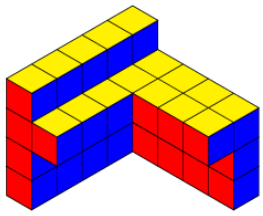
These results demonstrate that EQ3D inference generalises across dimensional configurations while preserving correctness. However, inferential completeness is inherently configuration-dependent: identical rule sets yield markedly different coverage levels depending on the distribution of visible depth information. This variability is not random but systematically linked to the balance between boundary-derived cues and interior visible structure. The mechanisms underlying this behavior are analysed in the following discussion.

Inference Output

All inferred matrices for the full set of twenty-four objects, including the six canonical views (FRU and BLD) and the cell-level inference outputs, are released as part of the open EQ3D dataset. For each object, the dataset contains: (i) the visible input views, (ii) the inferred hidden faces, (iii) the corresponding ground-truth matrices, and (iv) cell-level markers indicating determined and undetermined entries. A couple of examples can be seen in Figure 3.

This enables full reproducibility of all reported coverage values, independent verification of correctness, and detailed inspection of which cells are logically entailed by the visible projections and which remain underdetermined. The dataset makes explicit that undetermined cells are not inference errors, but geometric ambiguity under orthographic constraints.

o13



FRU

Front	Right	Up
$[*,*,*,*,*]$	$[*,*,*,*,*]$	$[b,c,c,c,c]$
$[a,*,*,*,*]$	$[e,e,e,e,e]$	$[b,c,c,c,c]$
$[a,a,d,d,d]$	$[d,d,d,a,a]$	$[b,c,*,*,*]$
$[a,e,e,e,e]$	$[e,e,e,e,a]$	$[b,c,*,*,*]$
$[a,e,e,e,e]$	$[e,e,e,e,a]$	$[b,c,*,*,*]$

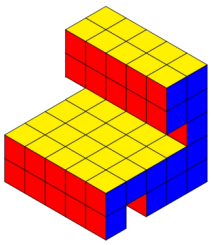
Inferred

Back	Left	Down
$[*,*,*,*,*]$	$[*,*,*,*,*]$	$[a,_,*,*,*]$
$[*,*,*,*,a]$	$[a,a,a,a,a]$	$[a,_,*,*,*]$
$[a,a,a,a,_]$	$[_,_,_,_,_]$	$[a,_,*,*,*]$
$[a,a,a,a,_]$	$[_,a,a,a,a]$	$[a,_,_,_,_]$
$[a,a,a,a,_]$	$[_,a,a,a,a]$	$[_,a,a,a,a]$

BLD (GT)

Back	Left	Down
$[*,*,*,*,*]$	$[*,*,*,*,*]$	$[a,c,*,*,*]$
$[*,*,*,*,a]$	$[a,a,a,a,a]$	$[a,c,*,*,*]$
$[a,a,a,a,a]$	$[a,a,a,a,a]$	$[a,c,*,*,*]$
$[a,a,a,a,a]$	$[a,a,a,a,a]$	$[a,c,c,c,c]$
$[a,a,a,a,a]$	$[a,a,a,a,a]$	$[a,a,a,a,a]$

o19



FRU

Front	Right	Up
$[d,d,d,d,d]$	$[*,*,*,a,a]$	$[a,a,a,a,a]$
$[d,d,d,d,d]$	$[*,*,*,a,a]$	$[a,a,a,a,a]$
$[e,e,e,e,e]$	$[*,*,*,*,a]$	$[d,d,d,d,d]$
$[a,a,a,a,a]$	$[a,a,a,a,a]$	$[d,d,d,d,d]$
$[a,a,a,a,a]$	$[a,*,a,a,a]$	$[d,d,d,d,d]$

Inferred

Back	Left	Down
$[a,a,a,a,a]$	$[a,a,*,*,*]$	$[a,a,a,a,a]$
$[a,_,_,_,_]$	$[_,_,*,*,*]$	$[_,_,_,_,_]$
$[a,a,a,a,a]$	$[_,*,*,*,*]$	$[_,_,_,_,a]$
$[a,_,_,_,_]$	$[_,_,a,a,a]$	$[_,_,_,_,a]$
$[a,_,_,_,_]$	$[_,_,_,*,_]$	$[_,_,_,_,a]$

BLD (GT)

Back	Left	Down
$[a,a,a,a,a]$	$[a,a,*,*,*]$	$[a,a,a,a,a]$
$[a,a,a,a,a]$	$[a,a,*,*,*]$	$[b,b,b,b,b]$
$[a,a,a,a,a]$	$[a,*,*,*,*]$	$[a,a,a,a,a]$
$[a,a,a,a,a]$	$[a,a,a,a,a]$	$[a,a,a,a,a]$
$[a,a,a,a,a]$	$[a,a,a,*,a]$	$[a,a,a,a,a]$

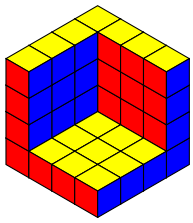
Figure 3 Inference comparison for objects o13 (76.00% coverage) and o19 (49.33% coverage).

6 Discussion

Scalability and Structural Generality

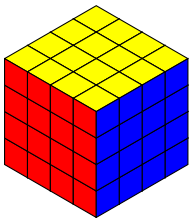
The experimental results confirm a central design claim of EQ3D: *logical soundness is invariant under dimensional scaling*. Across all tested configurations (from compact $2 \times 2 \times 2$ cubes to larger $6 \times 7 \times 6$ rectangular volumes) the inference rules required no modification, tuning, or size-specific adaptation. The rule system operates purely over symbolic adjacency relations, view correspondences, and ordinal depth constraints. As a consequence, scalability in EQ3D is *structural* rather than procedural: increasing grid resolution expands the domain of indices, but does not alter the reasoning mechanism itself.

At the same time, inferential *completeness* (measured through coverage) varies significantly across objects. While correctness remains constant (100% accuracy), coverage ranges from 26.85% (object 18) to 82.67% (object 9). This variability does not reflect instability; instead, it reveals that the proportion of determinable hidden structure is inherently dependent on qualitative configuration.



Front	Right	Up
$[a, d, d, d]$	$[d, d, d, a]$	$[a, a, a, a]$
$[a, d, d, d]$	$[d, d, d, a]$	$[a, d, d, d]$
$[a, d, d, d]$	$[d, d, d, a]$	$[a, d, d, d]$
$[a, a, a, a]$	$[a, a, a, a]$	$[a, d, d, d]$

(a) Boundary-driven extreme structure.



Front	Right	Up
$[a, a, a, a]$	$[a, a, a, a]$	$[a, a, a, a]$
$[a, a, a, a]$	$[a, a, a, a]$	$[a, a, a, a]$
$[a, a, a, a]$	$[a, a, a, a]$	$[a, a, a, a]$
$[a, a, a, a]$	$[a, a, a, a]$	$[a, a, a, a]$

(b) Uniform structure.

Figure 4 Conceptual contrast between two $4 \times 4 \times 4$ objects. (a) Depth extremes concentrated along inference boundaries with maximal depth (`d`) dominating interior cells. (b) Uniform shallow configuration with no depth variation. The first configuration provides stronger inferential constraints, while the second admits greater ambiguity in hidden-view reconstruction.

In other words, EQ3D scales deterministically across dimensions, but the *degree of determinacy* depends on how informative the visible structure is. Some objects provide strong symbolic constraints that tightly restrict hidden continuations; others admit multiple logically consistent completions. This distinction between *scalability of reasoning* and *configuration-dependent determinacy* is fundamental to understanding EQ3D's behaviour.

Concavity, Extremal Structure, and Inferential Reach

A key factor affecting coverage is the distribution of qualitative depth values across inference-relevant regions. EQ3D rules propagate information primarily from:

1. **Inference-boundary slices** (e.g., last column of `Right`, top row of `Up`, left column of `Up`, last row of `Right`, last row of `Front`), and
2. **Non-boundary cells** containing holes (`*`) or view-dependent maximum depth symbols.

Visible structure that exhibits strong depth contrasts at these locations tends to produce higher inferential reach. By contrast, configurations dominated by intermediate depth values in non-boundary regions leave more degrees of freedom for hidden views. This structural effect can be illustrated conceptually by contrasting two $4 \times 4 \times 4$ configurations (Figure 4).

In configuration (a), inference-boundary cells contain minimal depth (a), while interior cells contain maximal depth (d). This produces strong extremal cues: propagation across opposite views becomes tightly constrained, and fewer hidden alternatives remain admissible. Table 2 shows substantially higher coverage (66.67%) and a very low interior non-extremal ratio (INER = 0.10), indicating that most interior cells are structurally forced rather than ambiguous.

In configuration (b), uniform depth provides fewer constraints. Although globally consistent, the visible structure does not sufficiently restrict hidden extensions, resulting in lower inferential reach. This is reflected in Table 2 with the reduced coverage (37.50%) and maximal INER (1.00), meaning that all interior cells remain non-extremal and thus weakly informative despite the high boundary-rule share (BSS = 1.00).

Figure 3 shows this behaviour. Object o13 exhibits higher inferential completeness, with broader propagation of depth constraints and lower ambiguity in hidden cells. In contrast, o19 shows reduced coverage and more undetermined entries, reflecting weaker interior constraints and heavier reliance on boundary-driven inference, consistent with the coverage and structural indicators reported in.

Table 2 Coverage and inference indicators for conceptual objects (a) and (b).

Object	Coverage (%)	Collision Rate (CR)	Boundary Share (BSS)	INER
(a) Boundary-driven	66.67	0.3191	0.4255	0.1000
(b) Uniform	37.50	0.1000	1.0000	1.0000

Correlation Patterns and Structural Effects

Correlation analysis between coverage and structural indicators reveals a coherent pattern. The strongest negative correlation with coverage is observed for `V_INTERIOR_OTHER_R` (Spearman $\rho = -0.837$). This metric measures the proportion of visible non-boundary cells whose values are neither holes nor maximum depth. A high value indicates heterogeneous intermediate depth structure. The strong negative correlation confirms that such configurations reduce inferential determinacy. Conversely, coverage correlates positively with indicators of extremal qualitative structure:

- **Maximum-depth ratio (MDR)**: Spearman $\rho = 0.533$,
- **Visible maximum-depth count (V_MAX)**: Spearman $\rho = 0.638$,
- **Combined star/max presence (SAS)**: Spearman $\rho = 0.642$.

These results confirm that holes and maximum-depth symbols act as strong inferential anchors. They reduce the number of admissible hidden completions and increase reconstructability. Boundary Share (BSS) shows a moderate negative correlation with coverage (Spearman $\rho = -0.600$). This suggests that heavy reliance on inference-boundary rules is characteristic of more ambiguous configurations: when interior propagation is weak, the system proportionally depends more on boundary-driven inference. Overall, the correlation profile supports a unified structural interpretation:

Key observation. Coverage increases when visible structure contains strong extremal qualitative cues and decreases when interior regions exhibit intermediate depths that admit multiple consistent completions.

Implications for Qualitative Volumetric Reasoning

These findings highlight a central epistemic property of EQ3D: the framework does not optimise over possible completions, nor does it approximate hidden structure. Instead, it performs principled symbolic propagation grounded in qualitative constraints.

Variation in coverage therefore reflects *intrinsic geometric ambiguity*, not algorithmic deficiency. EQ3D distinguishes between (i) structure that is logically entailed by visible projections, and (ii) structure that remains underdetermined. This distinction is essential in qualitative spatial reasoning, where interpretability and constraint-based inference take precedence over maximal reconstruction.

Limitations. EQ3D operates over discrete qualitative symbols and assumes perfectly aligned orthographic views. It does not currently model uncertainty in the input, such as ambiguous depth assignments or perceptual noise. Depth qualitative values are treated ordinally rather than metrically, and symmetrical configurations may yield multiple qualitatively valid hidden completions.

7 Conclusion and Future Work

This paper presented the EQ3D, a dimension-agnostic extension of the qualitative 3D (Q3D) representation that generalises volumetric reasoning from a fixed $3 \times 3 \times 3$ grid to arbitrary $H \times W \times L$ partitions. EQ3D preserves the symbolic depth vocabulary and the core logical principles of Q3D – depth relativity, volumetric continuity, and view adjacency, while removing dimensional constraints. In doing so, it establishes a uniform reasoning framework capable of handling heterogeneous volumetric resolutions without altering the underlying rule system.

A central contribution of EQ3D lies in the integration of adjacent-view propagation rules with the opposite-view constraints. This combination enables additional deductions from locally shared boundaries and increases inferential reach without introducing procedural complexity. On the other hand, the reasoning mechanism remains fully declarative and size-independent: validation and inference both scale linearly with the number of view cells and hidden targets.

Empirical evaluation across twenty-four structurally diverse objects shows that EQ3D consistently reconstructs hidden views from three orthographic inputs while preserving global cross-view coherence. All inferred cells match the ground truth (100% accuracy), and variation in coverage reflects intrinsic geometric ambiguity rather than algorithmic weakness.

Beyond reconstruction performance, EQ3D provides a principled characterisation of *inferential determinacy*. The EQ3D does not approximate or optimise over candidate completions; instead, it explicitly distinguishes between hidden cells that are logically involved and those that remain underdetermined. This property is a strength of EQ3D as a spatial reasoning model rather than merely a reconstruction technique, since it shows robust explainability.

As future work we intend to explore diverse directions. On one hand, we intend to explore a probabilistic extension of EQ3D, for instance through a ProbLog-based [5] formulation, allowing uncertain or partially reliable input views to be represented with graded confidence. Such an extension would enable reasoning under perceptual noise while preserving symbolic traceability of inference. This would relax the current assumption that the `Front-Right-Up` descriptions are exact inputs. A complementary direction involves developing a neural-symbolic EQ3D approach using a neural front-end which might predict soft `Front-Right-Up` tables from images, point clouds, or depth maps, while EQ3D might enforce structural consistency through logical constraints. This would also provide a possible inference to less regular geometries, since curved, oblique, or non-axis-aligned objects could first be approximated through voxels or part-based abstractions before being translated into qualitative EQ3D descriptors. For that, consistency-aware loss functions might ensure that learned predictions respect adjacency and depth relations, yielding an interpretable hybrid pipeline. On the other hand, recent advances in 3D object detection architectures suggest that the voxel-based and part-aware aggregation mechanisms can effectively capture spatial structure while maintaining computational efficiency [33, 6]. Integrating such representations with symbolic consistency constraints might provide a bridge between geometric learning

and logical reasoning. Moreover, as point-voxel transformers model global and fine geometric structure [22], supporting constraint-aware view prediction, then active-view selection strategies may be explored, where the system might determine which additional orthographic projection would maximally reduce hidden ambiguity.

Taken together, these directions suggest a trajectory in which EQ3D evolves from a purely qualitative inference calculus into a probabilistic, learnable, and perception-aware reasoning module. By maintaining a strong symbolic core while enabling integration with statistical and geometric components, EQ3D offers a promising foundation for interpretable neuro-symbolic reasoning over uncertain inputs and less regular volumetric structures.

References

1 American Dental Association. Dental admission test (dat). https://www.ada.org/en/education/testing/exams/dental-admission-test-dat, 2025. Includes the Perceptual Ability Test (PAT) section.

2 Anthony G. Cohn and Jochen Renz. Chapter 13 qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 551–596. Elsevier, 2008. doi:10.1016/S1574-6526(07)03013-1.

3 DAT Prep. Dat–pat: Perceptual ability test. https://www.dat-prep.com/dat-pat-perceptual-ability, 2025. Accessed: 2026-02-24.

4 Artur d'Avila Garcez, Luis C. Lamb, and Dov M. Gabbay. Neural-symbolic cognitive reasoning. *Springer Handbook of Computational Intelligence*, 2015.

5 Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. Problog: a probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, IJCAI'07, pages 2468–2473, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL: https://dl.acm.org/doi/10.5555/1625275.1625673.

6 Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1201–1209, May 2021. doi:10.1609/aaai.v35i2.16207.

7 Frank Dylla, Jae Hee Lee, Till Mossakowski, Thomas Schneider, André van Delden, Jasper van de Ven, and Diedrich Wolter. A survey of qualitative spatial and temporal calculi: Algebraic and computational properties. *ACM Comput. Surv.*, 50(1):7:1–7:39, 2017. doi:10.1145/3038927.

8 Martí Ejarque and Zoe Falomir. Ejarque24/Extended-Q3D. Software (visited on 2026-07-20). URL: https://github.com/Ejarque24/Extended-Q3D.

9 Z. Falomir. Towards a qualitative descriptor for paper folding reasoning. In *Proc. of the 29th International Workshop on Qualitative Reasoning*, 2016. Co-located with IJCAI'2016 in New York, USA.

10 Zoe FAlomir. A Qualitative Model for Reasoning about 3D Objects using Depth and Different Perspectives. In Tomasz Lechowski, Przemyslaw Andrzej Walega, and Michal Zawidzki, editors, *Proceedings of the 2015 Workshop on Logics for Qualitative Modelling and Reasoning (LQMR@FedCSIS 2015), Lódz, Poland, September 13-16, 2015*, volume 7 of *Annals of Computer Science and Information Systems*, pages 3–11, 2015. URL: https://annals-csis.org/Volume_7/drp/370.html.

11 Zoe Falomir. A qualitative model to reason about object rotations (QOR) applied to solve the cube comparison test (CCT). *ArXiv preprint arXiv.2601.08382*, 2025. doi:10.48550/arXiv.2601.08382.

12 Zoe Falomir, Albert Pich, and Vicent Costa. Spatial reasoning about qualitative shape compositions. *Ann. Math. Artif. Intell.*, 88(5-6):589–621, 2020. doi:10.1007/S10472-019-09637-7.

13 Zoe Falomir, Ruben Tarin, Aurelio Puerta, and Pablo Garcia-Segarra. An interactive game for training reasoning about paper folding. *Multim. Tools Appl.*, 80(5):6535–6566, 2021. doi:10.1007/s11042-020-09830-5.

14 Kenneth D. Forbus. Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(4):374–391, 2011. doi:10.1002/wcs.115.

15 Christian Freksa. *Qualitative Spatial Reasoning*, pages 361–372. Springer Netherlands, Dordrecht, 1991. doi:10.1007/978-94-011-2606-9_20.

16 Jinn-Kwei Guo, Chin-Hsing Chen, and Jiann-Der Lee. Multi-polyhedron reconstruction in a three-view system using relaxation. *Signal Processing*, 77(2):171–193, 1999. doi:10.1016/S0165-1684(99)00031-6.

17 Hamid Haghshenas Gorgani, Alireza Jahantigh, and Sadegh Sadeghi. 3d model reconstruction from two orthographic views using fuzzy surface analysis. *European Journal of Sustainable Development Research*, 3, March 2019. doi:10.29333/ejosdr/5726.

18 Wenyu Han, Siyuan Xiang, Chenhui Liu, Ruoyu Wang, and Chen Feng. Spare3d: A dataset for spatial reasoning on three-view line drawings. In *CVPR*, pages 14678–14687, June 2020. doi:10.1109/CVPR42600.2020.01470.

19 Long Hoang. 3d solid reconstruction from 2d orthographic views. In Branislav Sobota and Dragan Cvetković, editors, *Mixed Reality and Three-Dimensional Computer Graphics*, chapter 6. IntechOpen, London, 2020. doi:10.5772/intechopen.91977.

20 Wentao Hu, Jia Zheng, Zixin Zhang, Xiaojun Yuan, Jian Yin, and Zihan Zhou. Plankassembly: Robust 3d reconstruction from three orthographic views with learnt shape programs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18449–18459, 2023. doi:10.1109/ICCV51070.2023.01695.

21 David A. Huffman. Impossible objects as nonsense sentences. In *Machine Intelligence 6*, 2012.

22 Guangyu Ji, Jun Lu, Chengtao Cai, and Kaibin Qin. A point-voxel transformer for point cloud object detection with spatial and channel attention. *Engineering Applications of Artificial Intelligence*, 163:113141, 2026. doi:10.1016/j.engappai.2025.113141.

23 Lefteris M. Kirousis and Christos H. Papadimitriou. The complexity of recognizing polyhedral scenes. *Journal of Computer and System Sciences*, 37(1):14–38, 1988. doi:10.1016/0022-0000(88)90043-8.

24 Nikhil Krishnaswamy, Scott Friedman, and James Pustejovsky. Combining deep learning and qualitative spatial reasoning to learn complex structures from sparse examples with noise. In *Proc. of 33rd AAAI Conf. on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. doi:10.1609/aaai.v33i01.33012911.

25 G. Ligozat. *Qualitative Spatial and Temporal Reasoning*. MIT Press, Wiley-ISTE, London, 2011.

26 Jianzhuang Liu, Liangliang Cao, Zhenguo Li, and Xiaoou Tang. Plane-based optimization for 3d object reconstruction from single line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):315–327, 2008. doi:10.1109/TPAMI.2007.1172.

27 A. Lovett, M. Dehghani, and K. Forbus. Learning of qualitative descriptions for sketch recognition. In *Proc. 20th Int. Workshop on Qualitative Reasoning (QR), Hanover, USA, July*, 2006.

28 A. Lovett and K. Forbus. Cultural commonalities and differences in spatial problem-solving: A computational analysis. *Cognition*, 121(2):281–287, 2011. doi:10.1016/j.cognition.2011.06.012.

29 A. Lovett and K. Forbus. Modeling visual problem solving as analogical reasoning. *Psychological Review*, 124(1):60–90, 2017.

30 David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, July 2010. doi:10.7551/mitpress/9780262514620.001.0001.

31 Albert Pich and Zoe Falomir. Logical composition of qualitative shapes applied to solve spatial reasoning tests. *Cogn. Syst. Res.*, 52:82–102, 2018. doi:10.1016/J.COGSYS.2018.06.002.

32 Jochen Renz and Bernhard Nebel. On the complexity of qualitative spatial reasoning: a maximal tractable fragment of the region connection calculus. *Artif. Intell.*, 108(1–2):69–123, March 1999. doi:10.1016/S0004-3702(99)00002-8.

33 Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2647–2664, 2021. doi:10.1109/TPAMI.2020.2977026.

34 Sheryl Sorby. *Developing Spatial Thinking*. Higher Education Services, 2016.

35 Ken Tomiyama and Ken'ichi Nakaniwa. Reconstruction of 3D solid model from three orthographic views — Top-down approach. In Rangachar Kasturi and Karl Tombre, editors, *Graphics Recognition Methods and Applications*, pages 260–269, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

36 Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. SWI-Prolog. *Theory and Practice of Logic Programming (TPLP)*, 12(1-2):67–96, 2012. doi:10.1017/S1471068411000494.

37 Falomir Zoe and Oliver Eric. Q3D-Game: A Tool for Training Users' 3D Spatial Skills. In Paulo Novais and Shin'ichi Konomi, editors, *Intelligent Environments 2016 - Workshop Proc. of 12th Int. Conference on Intelligent Environments, IE 2016, London, UK, September 14-16, 2016*, volume 21 of *Ambient Intelligence and Smart Environments*, pages 187–196. IOS Press, 2016. doi:10.3233/978-1-61499-690-3-187.

38 Changqing Zou, Shifeng Chen, Hongbo Fu, and Jianzhuang Liu. Progressive 3d reconstruction of planar-faced manifold objects with drf-based line drawing decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):252–263, 2015. doi:10.1109/TVCG.2014.2354039.

On Spatial Reasoning and Perspective Transformation in Language Models

Haotong Zhang[1]
Department of Computer Science, University of Manchester, UK

Ian Pratt-Hartmann
Department of Computer Science, University of Manchester, UK
Instytut Informatyki, Uniwersytet Opolski, Opole, Poland

Abstract

In this paper, we investigate the performance of language models on spatial reasoning with textually presented data, with particular focus on the ability to switch between different perspectives (route vs. survey) – an important component of human spatial reasoning. Our methodology involves the construction of large, automatically labelled corpora, as opposed to crowd-sourced, human-annotated datasets; this approach focuses attention on the language models' command of underlying geometrical principles, rather than their access to background knowledge and commonsense rules-of-thumb. Results reveal that language models can acquire basic spatial reasoning ability after finetuning on targeted tasks, though they do not fully capture the underlying rules. Furthermore, language models show significant differences from humans in the perspective-transformation task, exhibiting distinct patterns of performance across perspective conditions.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Spatial Reasoning, Language Models, Perspective Transformation, Spatial Cognition

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.6

Supplementary Material *Software (Source code and datasets)*: `https://github.com/zhanghaotong1/cosit-2026-spatial-reasoning-and-perspective-transformation-in-LMs`
archived at `swh:1:dir:c9382b808e5a17a051732b9787ceacd5e16c3c2c`

Funding *Haotong Zhang*: Supported by the University of Manchester.
Ian Pratt-Hartmann: Supported by the Polish NCN, grant number 2025/57/B/ST6/04170.

Acknowledgements We would like to acknowledge the use of the Computational Shared Facility at the University of Manchester. We also thank the anonymous reviewers for their helpful feedback.

1 Introduction

By *spatial reasoning*, we understand the faculty of making inferences mediated by the reasoner's knowledge – implicit or explicit – of the space in which it is operating. This definition comprises all types of spatial knowledge from the banal (if you are heading due north, objects to the east will appear on your right) to the recherché (rotating an object through 180° about each of three orthogonal axes will return it to its original orientation). Spatial reasoning underlies the ability to execute numerous everyday (as well as some not-so-everyday) tasks. Such is its centrality in human – and indeed animal – intelligence, that it has generated a great quantity of psychological research. How do humans (or other animals) solve spatial reasoning problems? What representations do they employ to describe spatial

[1] Corresponding author

© Haotong Zhang and Ian Pratt-Hartmann;
licensed under Creative Commons License CC-BY 4.0

arrangements? What algorithms do they use to manipulate them [4]? At the same time, the astonishing development of transformer-based language models (LMs) and their more recent evolution into large *reasoning* models [64] have demonstrated at least an apparent ability on the part of these architectures to engage in humanlike reasoning, suggesting some grasp of rudimentary logical principles. The question therefore arises as to how well such models can engage in *spatial* reasoning, and, more specifically, what their characteristics are [6]. Can LMs reason about space? If so, do they reason in ways that resemble the patterns observed in humans? That is the subject of this paper.

When considering research into the reasoning abilities of LMs, we must distinguish two very different paradigms. The first understands inference as a matter of arriving at *probably correct* judgments mediated by the sort of commonsense, background knowledge expected of every normal person [3, 44, 63]. For example, from *The cat is in the garden*, it may be inferred that *The cat is not in the kitchen*, using the background knowledge that gardens are seldom found in kitchens (and that, quantum effects aside, cats cannot be in two places at once). The second paradigm examines entailment *sensu stricto* [48, 46, 56, 17]: a collection of premises Φ is said to entail a conclusion ψ if all logically possible situations making all the premises Φ true also make ψ true. For example, from the two premises *All artists are creative* and *Some beekeepers are artists*, we can infer *Some beekeepers are creative* as a matter of logic. These two paradigms differ sharply in the datasets typically employed. Whereas the former pradigm is largely reliant on human-annotated, crowd-sourced data, the latter is free to exploit automatically generated corpora of essentially unlimited size. For example, in training systems to recognise logically valid inferences couched in natural language, we may automatically construct large collections of premises and (putative) conclusions in fragments of language specified by some grammar, and label them (valid or invalid) using an automated theorem prover. As for reasoning in general, so for spatial reasoning in particular: corpora of reasoning problems – even those couched in natural language – can be constructed and labelled using fully automated techniques [7]. It is this second paradigm that we adopt here.

The problem of determining logical entailments mediated by knowledge of the operative space is, of course, the domain of geometry, broadly construed. Where considerations of the language employed and the algorithms required to process them are to the fore, we are in the more specific realm of *spatial logic*. By a spatial logic, we understand a formal language whose variables range over some prescribed domain of geometrical entities – *e.g.*, points, lines, regions of space – and whose non-logical primitives are interpreted as relations defined on those geometrical entities – *e.g.* incidence, parallelism, inclusion. The relations considered may be purely qualitative (as in topological calculi, or indeed as in traditional presentations of Euclidean geometry) or quantitative (as in real algebraic geometry). For a survey of spatial logic, see [1]. When investigating the performance of LMs on a particular class of spatial reasoning problems, the mathematical and computational aspects of the class of problems in question must be thoroughly understood. Thus, we find ourselves at the confluence of three research programmes – psychological studies of human spatial reasoning, computational investigations of spatial logic, and the use of LMs from artificial intelligence. What spatial inferences can LMs perform effectively? How does this performance correlate with the intrinsic complexity of the problems in question? Do these models exhibit similar behaviour to humans in respect of spatial reasoning?

The present paper tackles these issues from the perspective of a particular problem which has attracted the attention of cognitive psychologists, namely that of shifting between route-based and survey-based views. Route-based spatial descriptions adopt a first-person perspective: *As you leave the Capitol going along the Mall, the first building you pass on your*

right is ... Survey-based descriptions, by contrast, take the view from nowhere: *At the east end of the Mall stands the Capitol and at the west end ...* The (obviously useful) ability to switch between these sorts of descriptions – *perspective transformation* – has been the object of a certain amount of psychological study, notably by B. Tversky [57, 55, 58]. How, we ask, do LMs perfom on such tasks? To help focus our investigations, we proceed step-by-step, considering first a range of reasoning problems where the underlying spatial logic is well understood, and investigate the extent to which LMs can learn to solve these problems. Specifically, we define three spatial reasoning tasks – which we call the one-dimensional (**1D**) task, the two-dimensional (**2D**) task and the oriented matroid realisability (**OMR**) task – and construct datasets of sets of statements, couched in natural language – representing descriptions of spatial arrangements. Broadly speaking, we may regard the **2D**-task as involving reasoning within the *survey view*, while the **OMR**-task involves reasoning within the *route view*. The datasets in question are automatically labelled as either *satisfiable* (*i.e.* geometrically realisable) or *unsatisfiable* (geometrically unrealisable), and used to finetune pre-trained LMs to perform this classification. From an algorithmic perspective, the **1D**- and **2D**-tasks are relatively simple in their underlying logical structure, whereas the **OMR**-task involves more intricate reasoning. In each case, datasets are crafted to eliminate data artefacts. The difficulty of each problem instance is also measured to make sure that our datasets are non-trivial. With a preliminary understanding of the spatial reasoning abilities of LMs in place, we turn our attention to the issue of perspective transformation. Specifically, we define two further spatial reasoning tasks – which we call the **2D2OMR**-task, and the **OMR22D**-task. These tasks are intended – very approximately – to replicate perspective-transformation experiments reported in the psychological literature.

Throughout this paper, we use the term *spatial reasoning* in an operational sense. That is, we evaluate whether a model can correctly solve formally defined spatial inference tasks expressed in natural language. Our use of the term does not presuppose that language models employ reasoning processes identical to those used by humans, nor does successful task performance necessarily imply human-like spatial cognition. Rather, the tasks provide a behavioural framework for comparing the spatial inference capabilities of different systems.

2 Related Work

Spatial reasoning first caught the attention of psychologists in the last century, who carried out experiments to explore how humans comprehend spatial information, how they store such information as mental models and how they reason with mental models. Experiments revealed that humans construct mental models following certain processes and tend to overlook alternative possibilities [4, 43]. Another line of research focuses on perspective taking, especially between route (egocentric) and survey (allocentric) perspectives. Early work by Siegel and White [53] distinguished between route view and survey view in the development of spatial representations. Subsequent work further examined the respective roles of landmark, route, and survey knowledge in navigation [61]. Empirical studies have demonstrated that these two forms of representation differ not only in format but also in the cognitive processes required to construct and manipulate them [57, 55, 58]. Further studies indicate that original orientation also affects reasoning accuracy [49, 50].

It is instructive to consider many of these reasoning problems from a purely logical perspective. By a *spatial logic*, we understand a formal language interpreted over a class of structures featuring geometrical entities and relations [54]. A central line of research in spatial logic concerns qualitative spatial reasoning (QSR), which develops formal calculi for representing and reasoning about spatial relations without relying on precise metric

information [9]. Foundational work in this area includes topological calculi for regions [45], directional calculi for points [27, 28], and east and west calculi for intervals [26]. Another set of approaches is often motivated by observer-dependent spatial descriptions and their relationship to human spatial cognition. Examples include acceptance regions for projective spatial expressions, which have been argued to better reflect human usage of relative directional concepts [36]. Related approaches include the Dipole Calculus [37], which represents spatial entities as oriented line segments and supports reasoning about intrinsic orientation, and the OPRA family of calculi [38, 39], which models relative directions between oriented points at configurable granularity levels. These formalisms provide cognitively motivated representations of relative direction and have been applied to navigation and qualitative spatial reasoning tasks. A salient issue in QSR involves determining the computational complexity of particular spatial logics, especially with regard to the satisfiability-checking problem [21, 13, 11]. Of particular interest here are relatively expressive spatial logics whose satisfiability problems are decidable, ideally with low complexity.

In the last decade, textual spatial reasoning has been introduced to the field of NLP. BAbI was the first dataset to include spatial reasoning [62]. It was relatively simple and has since been largely solved by neural models. Later, three more complex datasets with more relations, entities and longer reasoning paths were introduced: SpartQA [35], SpaRTUN [34] and StepGame [51]. StepGame has been influential as a synthetic benchmark for multi-step textual spatial reasoning, although later work has identified template errors and inconsistencies in its definition of reasoning hops [25, 24]. These datasets nevertheless proved to be difficult for LMs such as BERT [10], ALBERT [22] and XLNet [68]. Premsri and Kordjamshidi claimed that the unsatisfactory performance of LMs on spatial reasoning is partly due to their poor generalisation ability on unseen data samples [41]. Therefore, they proposed training LMs with neuro-symbolic techniques. Although proven to be effective, the approach did not address how well LMs can perform on purely textual spatial reasoning.

With the emergence of large language models (LLMs), their performance on spatial reasoning in few-shot or zero-shot settings (*i.e.*, without finetuning) has attracted much attention [20, 5, 29, 8]. New benchmarks have been constructed and tested on LLMs [66, 47]. Models are reported to exhibit shortcomings on a variety of spatial reasoning benchmarks [2, 25, 65, 67]. Nevertheless, Li, Hogg, and Cohn [25] showed that, on a corrected version of the StepGame benchmark, LLMs can accurately extract spatial relations from natural-language descriptions, and that their performance can be improved through Chain-of-Thought prompting [60] and Tree-of-Thoughts prompting [69]. More recently, neuro-symbolic approaches have also been explored for spatial reasoning benchmarks similar to StepGame [33]. Furthermore, LLMs have been reported to exhibit certain performance patterns and biases similar to those observed in humans [6, 14, 7].

Recent work has examined spatial reasoning and perspective-taking in vision-language models (VLMs) and multimodal large language models (MLLMs), typically focusing on viewpoint transformations between different observers. Empirical analyses report systematic limitations in spatial inference in MLLMs [52]. Visual perspective-taking has been evaluated through tasks requiring models to adopt alternative observer viewpoints, revealing persistent difficulties in shifting away from egocentric representations [16]. More recent studies further document strong egocentric biases in VLMs [59], and suggest that incorporating cognitively inspired mechanisms can improve performance on cross-perspective reasoning tasks [23]. While related, these studies primarily address observer-based perspective shifts. In contrast, our work focuses on transformations between route and survey perspectives in linguistic spatial reasoning, which have been central in research on human spatial cognition and remain comparatively underexplored in studies of LMs.

In summary, our study links human perspective transformation, qualitative spatial logic, and natural language processing through a logical framework for spatial reasoning.

3 Reasoning Tasks

In this section, we define our spatial reasoning tasks. We first begin with a very simple task, which deals with points in one-dimensional space. A statement in this task takes the form $A\ r\ B$, interpreted as *A is on the r of B*. A and B represent points and r represents the spatial relation between A and B: left, right or overlap. Left and right follow our daily convention, while overlap means that two points are at exactly the same position. The task is formatted as a satisfiability-checking problem. From the geometric point of view, a set of statements is *satisfiable* just in case a set of points can be arranged on a line in accordance with the statements. We call this task the **1D**-task.

Next, we define a similar task which is only slightly more complicated than the **1D**-task: the **2D**-task. As indicated by its name, we now consider points in two-dimensional space. Statements are again of the form $A\ r\ B$, interpreted as *A is to the r of B*, except that r represents cardinal and intercardinal directions: north, south, east, west, northeast, southeast, northwest, southwest and overlap. This set of directional relations is closely related to the projection-based notion of cardinal directions introduced by Frank [12], and more broadly to the Cardinal Direction Calculus (CDC) literature. When we refer to "north", we specifically mean due north (*i.e.*, 0° on the compass). And "northeast" means any direction strictly between due north and due east. The same convention applies to other (inter)cardinal directions. We say a set of statements is satisfiable if there is a set of points in the two-dimensional space satisfying all statements at the same time.

The third task is the oriented matroid realisability problem (or **OMR** for simplicity). In this task, locations are described from an egocentric point of view. A problem instance consists of a set of statements having either of the forms $\mathit{left}(A, B, C)$ or $\mathit{right}(A, B, C)$, interpreted as: *standing at A and facing B, C is on the left/right*. Here A, B, C are objects that may regarded as unextended (*i.e.*, single points). We do not differentiate between front and back. The term "left" encompasses both front-left and back-left regions, and "right" is defined in an analogous manner. Note that "left" and "right" are interpreted in the strict sense. Thus, $\mathit{left}(A, B, C)$ or $\mathit{right}(A, B, C)$ is satisfied only when C lies strictly in the corresponding half-plane. Collinear configurations, where A, B, and C lie on the same line, are excluded and satisfy neither relation. Furthermore, the points A, B, C must be distinct; coincident points are not permitted. Despite the statements' simple form, this task is more complicated than the above two tasks, as each statement involves the relative orientation among three points. A set of statements is satisfiable if we can find a set of points on the two-dimensional Euclidean plane making all statements true. For example, $\mathit{left}(A, B, C), \mathit{left}(A, D, B), \mathit{left}(A, C, D), \mathit{left}(B, C, D)$ together is satisfiable (see Figure 1 for a realising arrangement). However, a closer look reveals that changing the last statement to $\mathit{left}(B, D, C)$ leads to unsatisfiability, no matter how we arrange A, B, C and D.

Finally we introduce two perspective transformation tasks, which are the main focus of this paper. Notice that our **2D**-task is similar to the survey view, in the sense that each statement describes the absolute direction between two points. On the other hand, our **OMR**-task can be regarded as the route view. Previous cognitive psychology experiments conducted on humans are formulated as question-answering tasks, where subjects are given a textual description of an area in one perspective and asked to determine whether a question from the other perspective is true or false. However, to maintain consistency with above tasks,

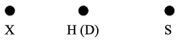
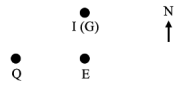
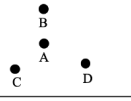
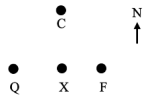
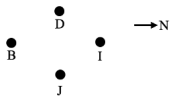
Task	Statements	Natural Language Form	Realising Arrangement
1D	$H\ left\ S$	H is on the left of S.	
	$H\ overlap\ D$	H overlaps with D.	
	$X\ left\ H$	X is on the left of H.	
	$D\ right\ X$	D is on the right of X.	
2D	$I\ overlap\ G$	I overlaps with G.	
	$E\ south\ I$	E is to the south of I.	
	$I\ northeast\ Q$	I is to the northeast of Q.	
	$E\ east\ Q$	E is to the east of Q.	
OMR	$left(A,B,C)$	Standing at A and facing B, C is on the left.	
	$left(A,D,B)$	Standing at A and facing D, B is on the left.	
	$left(A,C,D)$	Standing at A and facing C, D is on the left.	
	$left(B,C,D)$	Standing at B and facing C, D is on the left.	
2D2OMR	$C\ north\ X$	C is to the north of X.	
	$C\ northeast\ Q$	C is to the northeast of Q.	
	$F\ southeast\ C$	F is to the southeast of C.	
	$X\ east\ Q$	X is to the east of Q.	
	$left(C,X,F)$	Standing at C and facing X, F is on the left.	
OMR22D	$right(J,D,I)$	Standing at J and facing D, I is on the right.	
	$right(J,B,D)$	Standing at J and facing B, D is on the right.	
	$right(I,J,D)$	Standing at I and facing J, D is on the right.	
	$D\ northwest\ B \rightarrow I\ northeast\ D$	If D is to the northwest of B, then I is to the northeast of D.	

Figure 1 A satisfiable example in each task. The *Natural Language Form* column shows how to interpret logical statements into natural language and is the actual input to the models. We also give a realising arrangement for each example.

we again formulated two tasks as the satisfiability-checking problem. Specifically, in our **2D2OMR**-task, we append a single statement from the **OMR**-task (*i.e.* $left/right(A,B,C)$) to the end of a set of statements from the **2D**-task (*i.e.*, $A\ r\ B$), switching from the survey view to the route view. This setting is roughly similar to the psychological experiments in the sense that LMs first see a set of statements in the survey view (textual description), then a statement in the route view (question). Checking whether the question statement is consistent with the textual description amounts to checking whether they are jointly satisfiable. Similarly, in our **OMR22D**-task, we append a single statement from the **2D**-task to the end of a set of statements from the **OMR**-task[2]. This task requires the ability to switch from the route view to the survey view. In both tasks, a set of statements is satisfiable if a set of points satisfied all spatial relations described by the statements. While our setup is not a strict replication of previous psychology experiments, it captures the core feature of cross-perspective transformation between survey and route perspectives.

Figure 1 depicts a satisfiable example for each task. We also provide a realising arrangement alongside, which, as can be checked, satisfies every statement.

4 Data Construction

As mentioned in Section 1, datasets in our tasks can be labelled by automated algorithms, due to their purely logical character. An *instance* in the dataset comprises a set of statements. For the **1D**-task, it is straightforward that unsatisfiability can only rise from a *strict directed cycle*, namely a subset of statements of the form $A_1\ r_1\ A_2, A_2\ r_2\ A_3, ..., A_n\ r_n\ A_1$, where

[2] We also need to assign a random (inter)cardinal direction between two points since a pure route perspective does not contain any absolute direction information.

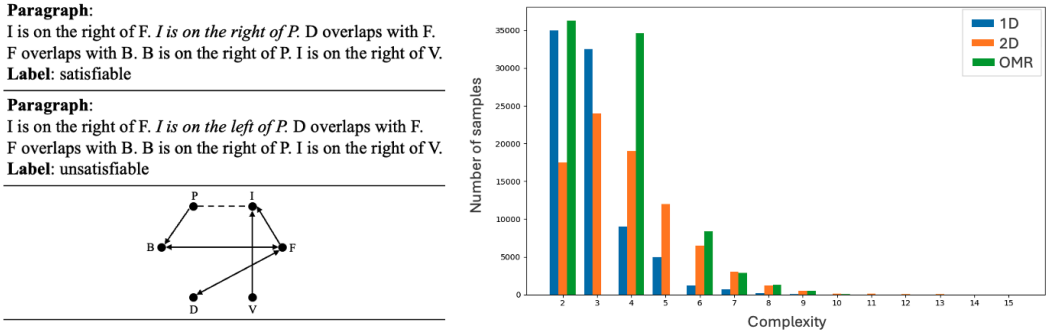


Figure 2 (a) A minimally distinguished pair from the **1D**-task. The second sentence (in *Italics*) is different in the two instances. To give a clearer illustration, we visualise the instances in a "graph-like" style where each point is a node in the graph, instead of plotting an one-dimensional arrangement. Directed edges (with one arrow) denote the left or reversed right relation. Bidirectional edges denote the overlap relation. The dashed edge between P and I corresponds to the different statement. If it goes from I to P (*i.e.*, I *left* P, as in the unsatisfiable instance), then there is a strict directed cycle $F - I - P - B - F$; (b) Complexity distribution in the **1D**-, **2D**- and **OMR**-task. We omit the complexity distribution for the **2D2OMR**- and **OMR22D**-task due to computational infeasibility.

either: (i) each r_i is *left* or *overlap*, with at least one of them being *left*; or (ii) each r_i is *right* or *overlap*, with at least one of them being *right*. (Here, of course, we are implicitly taking A *right* B to be equivalent to B *left* A.) Labelling data in the **1D**-task is equivalent to detecting strict directed cycles in the statements.

We can label data in the **2D**-task in a similar way, except that now we need to check for two independent axes: north-south and west-east. For example, when evaluating the north-south axis, the east and west direction is ignored: A is considered "latitudinally north" of B if it lies to the north, northeast or northwest of B. The same convention applies when evaluating the east–west axis: A is considered "longitudinally east" of B if it lies to the east, northeast or southeast of B. We remark that, here, we imagine points located in a small region, and disallow circumnavigation of the globe: thus, Berlin is north*east* of Paris, not north*west*.) A set of statements is easily seen to be satisfiable if there are no strict directed cycles along either axis.

Next we turn to the **OMR**-task. Checking satisfiability is equivalent to the problem of solving a set of quadratic inequalities in the real domain. Specifically, a statement $left(A, B, C)$ holds if and only if $(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1) > 0$, where $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ are coordinates of A, B, C in the two-dimensional Euclidean plane, respectively. We refer readers to, for example, [40], for a detailed proof. In principle, the feasibility of a quadratic inequality system can be determined exactly using symbolic solvers such as Mathematica[3]. However, exact solving becomes computationally expensive when the number of points grows. In constructing the dataset, we therefore rely on iterative quasi-Newton methods implemented in SciPy library[4] in Python to obtain approximate solutions. These methods provide solutions that satisfy the constraints up to a numerical tolerance (we use 1e-6) rather than guaranteeing exact symbolic correctness, but offer substantially improved efficiency and scalability for large numbers of instances. We randomly check some instances using Mathematica to make sure that SciPy is highly accurate.

[3] `https://www.wolfram.com/mathematica`
[4] `https://scipy.org`

We also use SciPy to label data in our perspective transformation tasks. As mentioned in Section 3, a single statement from one task is appended to the end of a set of statements from the other task (*i.e.*, perspective). We deliberately make sure that statements from the same perspective are satisfiable, to prevent intrinsic unsatisfiability from arising within one perspective. In other words, unsatisfiability can only come from the inconsistency during perspective transformation.

To eliminate potential shortcuts in instances that are easily detected by LMs and reduce task difficulty, we generate satisfiable and unsatisfiable instances in pairs for all tasks. Specifically, we first randomly generate sets of statements until we find an unsatisfiable one. Then we randomly reverse a statement in the set and check if the new set of statements is satisfiable. By "reverse", we mean to replace the spatial relation with its inverse relation[5] (*e.g.*, replace "left" with "right", replace "north" with "south"). If so, we call these two instances a *minimally distinguished pair*, and say that they are *contra-instances* of each other. We generate datasets in this way so that satisfiable and unsatisfiable instances cannot be easily distinguished by any superficial characteristics. In fact, two instances in a minimally distinguished pair only differ by one word. Figure 2 (a) shows an example of such pair in the **1D**-task and embodies the strict directed cycle in the unsatisfiable instance.

For an instance, both the number of statements (*instance size*) and the number of points involved (*domain size*) influence the probability of satisfiability of a random instance. Through some pilot experiments, we empirically show that setting both to be equal makes it easy to generate unsatisfiable instances while allowing satisfiable contra-instances to be found for most of them. The dataset construction process is as follows: (1) Generate a set of statements with determined instance size and domain size. (2) If the generated instance is unsatisfiable and there exists a satisfiable contra-instance, we add both to the dataset. (3) Repeat (1) and (2) until we have enough instances. Notice that the dataset constructed in this way is automatically balanced. We obtain a dataset for each of our tasks and randomly split each into a training set with around 74K instances and a test set with 10K instances. We make sure that each minimally distinguished pair is in the same set to avoid potential data leakage. Domain size varies from 3 to 26 (inclusive), with an equal number of instances for each domain size. Spatial relations are sampled uniformly during dataset generation. As a result, all relation types occur with approximately equal frequency in the datasets.

We additionally checked for duplicate instances and ensured that no identical instance appears in both the training and test sets. However, because points are represented by arbitrary uppercase letters, there may still exist pairs of instances that become identical after a permutation of uppercase letters. For example, (1) *T is on the left of K. K is on the left of B. B is on the left of T.* and (2) *U is on the left of I. I is on the left of Z. Z is on the left of U.* are essentially the same under uppercase letter permutation. Such instances are not treated as duplicates by our generation procedure. Although this does not affect the correctness of the dataset, it may reduce the effective diversity of the data and could potentially be exploited by models that learn to ignore specific point names. We leave a more systematic analysis of this issue for future work.

Finally, we provide a complexity measurement for an instance, to help use assess the difficulty of our datasets. Intuitively, a contradiction in an unsatisfiable instance is generally harder to detect if it is derived from more statements. Following this idea, we define the *complexity* of an unsatisfiable instance by the size of the *minimal unsatisfiable subset*, which is the smallest subset of statements that leads to a contradiction. For example, the instance

[5] For instances in the **2D2OMR**- or **OMR22D**-task, we only reverse the last (appended) statement.

A *left* B, B *left* C, C *left* A, A *right* D has complexity 3 since the first three statements form a strict directed cycle and any two statements do not contradict each other. Similarly, the instance $\mathit{left}(A, B, C), \mathit{left}(A, D, B), \mathit{left}(A, C, D), \mathit{left}(B, D, C)$ has complexity 4 because four statements together are unsatisfiable while any 3 statements are satisfiable. The complexity of a satisfiable instance is defined as the complexity of its contra-instance for simplicity. Figure 2 (b) shows the complexity distribution for the **1D**-, **2D**- and **OMR**-task in our constructed datasets, with domain size ranging from 3 to 26.

5 Experimental Setup

5.1 Language Models

To explore how well LMs can solve the satisfiability-checking problem in our tasks, we finetune three (small) language models, namely, RoBERTa large, DeBERTa-v3 large and T5 base[6]. This choice allows us to study spatial reasoning under relatively controlled conditions and to more clearly separate task learning from the complex and often opaque effects of large-scale pretraining. Our goal is therefore not to benchmark state-of-the-art large language models (LLMs), but to investigate how spatial reasoning can be acquired and represented in language models. To provide a point of comparison with modern LLMs, we additionally evaluate Gemini-3.5-Flash [15] in a zero-shot setting for the main experimental comparisons.

RoBERTa. The Robustly optimized BERT approach (RoBERTa) is trained in a more optimised manner based on the BERT architecture [30]. It removes the next sentence prediction objective, trains on longer sequences over more data and changes the masking pattern applied to the training data. It outperforms BERT on GLUE, SQuAD and RACE. We use RoBERTa large with 355M parameters.

DeBERTa-v3. The new architecture DeBERTa improves the BERT and RoBERTa models using a disentangled attention mechanism and an enhanced mask decoder [19]. The DeBERTa model consistently performs better than BERT and RoBERTa. DeBERTa-v3 further improves the efficiency of DeBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing [18]. We use DeBERTa-v3 large with 304M parameters.

T5. The Text-to-Text Transfer Transformer (T5) reframes all NLP tasks into a unified text-to-text-format where the input and output are text strings [42]. It is one of the state-of-the-art encoder-decoder language models. We employ T5 base with 220M parameters.

For all models, we utilise the Hugging Face[7] implementation with a classification layer on the top.

5.2 Implementation Details

Formally, we define the satisfiability-checking problem as a binary classification task. The input to the model is an instance (*i.e.*, a set of statements) that describes spatial relations between points, couched in natural language. The model needs to predict satisfiability of the input. Since the dataset is balanced, we use accuracy as the evaluation metric. For Gemini,

[6] We also considered T5 large, but it was unstable on our tasks, with several runs failing to converge.
[7] `https://huggingface.co`

Table 1 Accuracy of four models on our reasoning tasks. The first three models were finetuned on our task. Cells in the first three rows report post-finetuning (mean accuracy and standard deviation over three runs with random seeds 42, 418, and 2048) and pre-finetuning accuracies, shown before and after the slash (/), respectively. Gemini was evaluated without task-specific finetuning; we only report its zero-shot performance. *Model failed to converge even with smaller learning rates.

	1D	2D	OMR
RoBERTa large	99.03 ± 0.39 / 50.43	76.79 ± 2.78 / 50.12	50.03 ± 0.06* / 50.69
DeBERTa-v3 large	99.99 ± 0.01 / 51.05	99.42 ± 0.16 / 51.40	91.99 ± 0.02 / 50.06
T5 base	93.06 ± 3.22 / 50.26	67.69 ± 2.77 / 50.28	50.20 ± 0.16* / 50.05
Gemini-3.5	85.17	92.20	84.80

performance is evaluated in a zero-shot setting with standard prompts (Appendix A). We use the official API with temperature set to 0 to ensure deterministic outputs. Results are reported from a single run due to API cost constraints.

During finetuning, we optimise the standard cross-entropy loss for binary classification. We use the AdamW optimizer [31] with default parameters except for the learning rate. The learning rate is linearly warmed up from 0 to 1e-6 (for RoBERTa large and DeBERTa-v3 large) or 1e-5 (for T5 base) during the first epoch, followed by a linear decay to 0. We use a batch size of 24 and finetune each model for 10 epochs. These hyperparameters may not be optimal for every model; however, since the models already achieve strong and consistent performance, we do not conduct extensive hyperparameter tuning, accepting a trade-off between model-specific optimisation and experimental comparability. During inference, we pick the highest logit as the model's prediction.

6 Results and Discussion

6.1 Overall Model Performance

All investigated LMs yielded satisfactory performance on the **1D** and **2D** tasks after finetuning. By contrast, before finetuning the models performed at approximately chance level (50% accuracy), as shown in Table 1. This demonstrates that finetuning enables LMs to distinguish between minimally contrasting pairs. Accuracy decreases from the **1D** task to the **2D** task and further to the **OMR** task, which accords with our intuition that these tasks exhibit progressively greater difficulty. The performance of DeBERTa-v3 on the computationally demanding OMR task is particularly impressive.

The difference between models' performance increases from the **1D**-task to the **2D**-task to the **OMR**-task. All models performed remarkably well on the **1D**-task, while some of them dropped significantly on the other two tasks, or even failed to converge. Among them, DeBERTa-v3 always performed the best, even better than RoBERTa, which has more parameters. This might be due to its more effective architecture.

As a point of comparison with a powerful general-purpose language model, we additionally evaluated Gemini-3.5-Flash in a zero-shot setting. Gemini-3.5 achieved performance comparable to that of the finetuned language models across all three tasks, although its accuracy remained below that of DeBERTa-v3, the best-performing model. The strong zero-shot performance indicates that contemporary state-of-the-art LLMs already capture a considerable amount of spatial reasoning knowledge during pretraining, allowing them to generalise to our tasks without additional task-specific supervision.

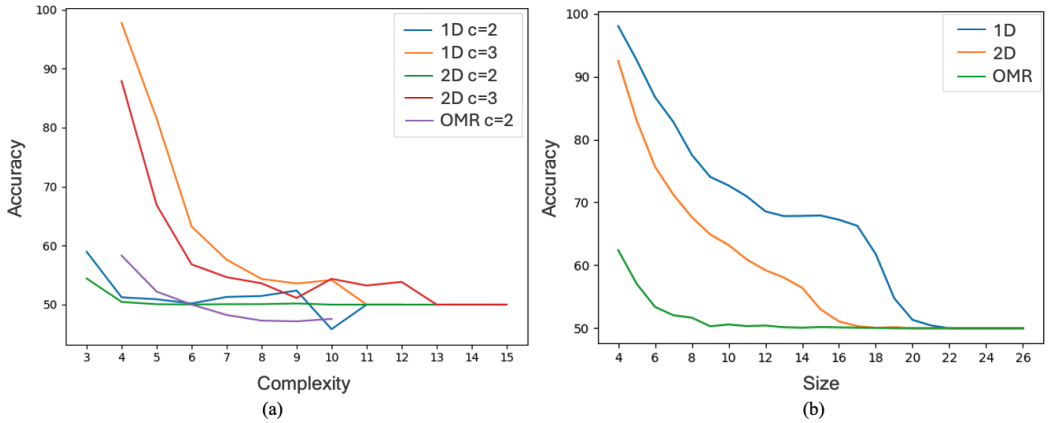


Figure 3 (a) Finetune DeBERTa-v3 on instances with complexity 2 or 3 and test on instances with higher complexity. c=2/3 indicates which complexity is the model finetuned on; (b) Finetune DeBERTa-v3 on instances with size 3 and test on instances with larger sizes.

6.2 Generalisation Ability

In the following experiments, we will use DeBERTa-v3 since it performed the best on all tasks. Figure 3 shows that DeBERTa-v3 exhibits different generalisation behaviour according to instance complexity or instance size. When the model is finetuned on instances with complexity 2, it shows poor generalisation ability on higher complexity, with accuracy only slightly better than a random guess (*i.e.*, 50%). Since complexity 2 means two statements directly contradict each other, only learning this type of contradiction barely helps the model to detect more difficult cases. By contrast, when the model is finetuned on instances with complexity 3, it is exposed to more complicated situations.[8] Therefore, it acquires the ability to handle more difficult instances to some extent. Nevertheless, accuracy quickly drops to below 70% above complexity 6. Confusion matrices show that the model tends to predict unsatisfiable instances as satisfiable, which reveals that the model fails to detect much longer strict directed cycles than what it has seen during finetuning.

When the model is finetuned on instances with instance size 3, and tested on larger instance sizes, it demonstrates better generalisation ability than when it is finetuned on instances with small complexity. A possible reason is that, although the instance size increases, the complexity distribution remains relatively similar, which is seen by the model during finetuning. Therefore, the model can utilise the learnt knowledge within a reasonable scope. Despite that, accuracy falls back to random guess when the instance size is very large. This is anticipated as a larger instance size means it is more difficult to find the statements that lead to contradiction. In other words, the model is not robust to "distractors" (*i.e.*, statements that are not in the minimal unsatisfiable subset).

In both cases, it is obvious that the model shows best generalisation ability on the **1D**-task, with the worst generalisation ability observed on the **OMR**-task. As observed in [32], LMs can learn to understand logically significant words and grammatical constructs more easily from simpler fragments than harder ones.

[8] No instance has complexity 3 in the **OMR**-task. In other words, any three statements are satisfiable whenever every pair of them is satisfiable.

Table 2 Accuracy of DeBERTa-v3 when finetuned on one task and tested on the other two tasks. The column headings denote which dataset is the model finetuned on. The row headings denote which dataset is the model tested on.

Test \ Finetune	1D	2D	OMR
1D	100.00	99.39	58.22
2D	59.89	99.77	55.84
OMR	48.70	49.02	95.67

Finally – essentially as a sanity check – we test how well the model can generalise to out-of-domain tasks. Specifically, DeBERTa-v3 is finetuned on one task, then tested on the other two tasks. Results in Table 2 support that the **1D**- and **2D**-task are similar to each other, but are different from the **OMR**-task. The model is robust to the change in relation names, as the model finetuned on the **2D**-task achieved very high accuracy when tested on the **1D**-task. However, the model finetuned on the **1D**-task only attained an adequate result on the **2D**-task, probably because it did not learn to check for both axes (*i.e.*, north-south and west-east) during finetuning. The model finetuned on the **OMR**-task presents generalisation ability on the **1D**- and **2D**-task to some extent. This indicates that the model finetuned on the hardest task still manages to gain some general spatial reasoning ability to employ on simpler tasks, despite the discrepancy in underlying rules.

Analysing confusion matrices gives us additional insights. We observed that if the model is finetuned on a simpler task and tested on a harder task, then it tends to predict unsatisfiable instances as satisfiable. By contrast, if the model is finetuned on a harder task and tested on a simpler task, then more satisfiable instances are predicted as unsatisfiable. This asymmetric generalisation likely stems from the difference in data structure among different tasks. Unsatisfiable instances in the simpler task tend to involve more explicit contradictions, whereas those in the harder task often depend on more complicated interactions. As a result, training on the simpler task may lead to a conservative unsatisfiability criterion that fails to generalise to the harder task, while training on the harder task may induce a more aggressive criterion that overgeneralises to the simpler task.

6.3 Perspective Transformation

The above results have demonstrated that LMs can gain basic spatial reasoning ability to some extent after finetuning on targeted tasks. We then move to perspective transformation. Switching between route-based view and survey-based view is an important ability which is – sometimes implicitly – involved in everyday life. There have long been studies in psychology that try to understand whether different perspectives lead to different mental representations [57, 55, 49, 50]. These studies have revealed that humans display asymmetrical behaviour when switching between the two perspectives. We are interested in whether LMs exhibit similar bias as humans.

We finetune DeBERTa-v3 on the two training sets corresponding to the **2D2OMR**- and **OMR22D**-task, respectively, and test on the test sets. The model achieved 94.44% accuracy on the **OMR22D**-task, while only 81.92% accuracy on the **2D2OMR**-task. The model appears to be more proficient at transforming from route perspective to survey perspective, in contrast to the tendency observed in human subjects, who performed better when transforming from survey perspective to route perspective. If we regard the **2D**- and the **OMR**-task as "single-perspective tasks" (*i.e.*, no perspective transformation is needed

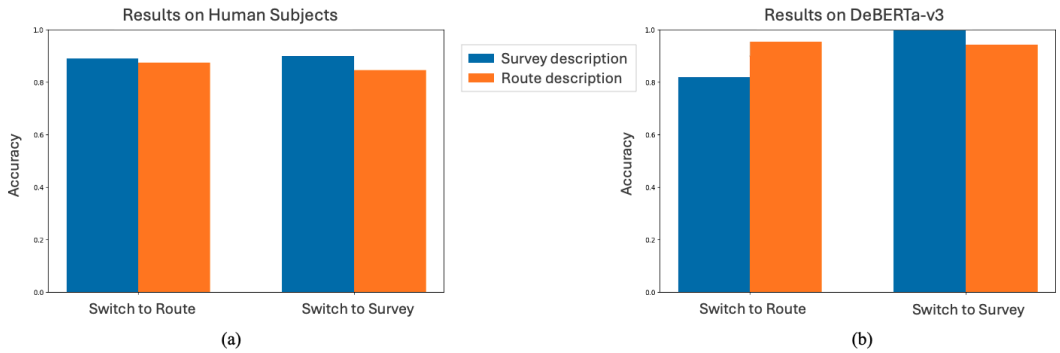


Figure 4 Performance of perspective transformation of human subjects (a) and DeBERTa-v3 (b). Survey/route description means the perspective taken in the textual description. x-axis indicates the perspective required to do inference. For humans, this is the perspective in which the question is asked. For the model, this is the perspective of the appended statement. Figure (a) is redrawn from Figure 5 in [57] based on the results reported in the paper. The figure is intended to illustrate qualitative performance patterns rather than provide a direct comparison of absolute accuracy.

during inference), comparing between these two tasks and the two "cross-perspective task" (*i.e.*, **2D2OMR**- and **OMR22D**-task) uncovers more differences between LMs and humans. Results on human subjects and DeBERTa-v3 are shown in Figure 4. Human results are drawn from a related experiment reported by Tversky [57] rather than being collected on our dataset. Moreover, our experimental setup is only inspired by, rather than a direct replication of, the original human study. In particular, the human experiment involved additional cognitive processes such as memory retention, delayed recall, and information retrieval, which cannot be straightforwardly mirrored in the language model setting. Accordingly, the comparison is intended to highlight qualitative similarities and differences in performance patterns across conditions, rather than to support a direct comparison of absolute accuracy.

Human performance appears to be more influenced by the perspective of the textual description itself: being given survey descriptions yields higher accuracy overall. In contrast, model performance seems more sensitive to whether perspective transformation is required, with single-perspective tasks generally outperforming cross-perspective tasks. As noted above, these comparisons should be interpreted cautiously, given the differences in datasets and experimental protocols. Nevertheless, the results suggest that humans and language models may be affected differently by perspective-related manipulations, leading to distinct patterns of performance across conditions.

Interestingly, Gemini-3.5 does not exhibit the same asymmetry. In the zero-shot setting, it achieved 85.77% accuracy on the **OMR22D**-task and 87.58% accuracy on the **2D2OMR**-task, comparable to the performance of finetuned DeBERTa-v3. This observation suggests that the asymmetry observed for DeBERTa-v3 may not generalise to all language models and warrants further investigation. At the same time, the strong performance of Gemini-3.5 suggests that contemporary LLMs can handle perspective transformations reasonably well even without task-specific supervision.

Finally, we investigate how the ratio of the number of statements from two perspectives influences model performance. We fix the total number of statements (equal to the number of points) in an instance to n, with the number of statements from the **2D**-task, $\#survey$, varying from 0 to n, and the number of statements from the **OMR**-task varying from n to 0. We choose n to be 5, 10, 15 and construct three balanced datasets. Each dataset is split into a training set with approximately 72K instances and a test set with approximately 10K instances. The number of instances with different ratios is the same within each dataset.

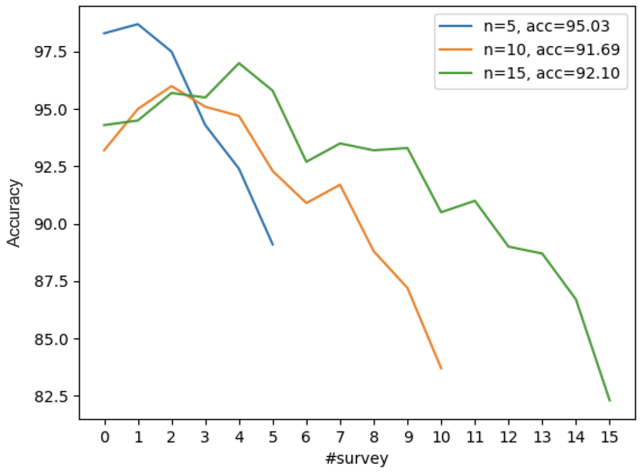


Figure 5 Accuracy breakdown based on the number of survey-perspective sentences on DeBERTa-v3. "acc" means the overall accuracy on the test set after finetuning.

We finetune DeBERTa-v3 on the training sets and test on the test sets. Figure 5 shows a trend of first increasing then decreasing with $\#survey$ increasing. The peak occurs when the proportion of survey-perspective statements (*i.e.*, $A\ r\ B$) is around 20%. We observe similar results as above that the model performed better when route-perspective statements (*i.e.*, $left/right(A, B, C)$) accounts for a larger proportion.

At present, we do not have a clear account of the observed results. While this pattern is robust across our experiments and is theoretically intriguing, we currently lack sufficient evidence to determine the underlying cause, and any explanation would therefore be highly speculative. We therefore wish to highlight this phenomenon as an important open question raised by the present study. Understanding why this specific mixture yields optimal performance may provide new insights into how language models integrate route-based and survey-based spatial information.

7 Discussion

Our findings have several implications for both the study of language models and the broader literature on spatial cognition. First, the results suggest that spatial reasoning should not be viewed as a single, homogeneous capability. Across our experiments, model performance varied substantially among the **1D**-, **2D**-, and **OMR**-tasks, despite all three being formulated as satisfiability-checking problems involving spatial relations. This observation indicates that different forms of spatial reasoning place qualitatively different computational demands on language models. In particular, the comparatively strong performance of DeBERTa-v3 on the **OMR**-task suggests that contemporary pretrained models can learn non-trivial forms of spatial-logical reasoning when provided with sufficient task-specific supervision.

Second, our perspective-transformation experiments suggest that reasoning within a perspective and transforming between perspective may constitute distinct challenges. Although DeBERTa-v3 achieved high accuracy on both the **2D**- and **OMR**-tasks individually, performance decreased when information had to be integrated across route and survey perspectives. This finding echoes long-standing observations in spatial cognition that different representational formats can place different demands on reasoning processes. More generally, it suggests that success on spatial reasoning tasks does not necessarily imply equal proficiency in perspective transformation.

A further implication concerns the relationship between language models and human spatial cognition. While our comparison with human results should be interpreted cautiously due to differences in datasets and experimental protocols, the observed performance patterns do not appear to align well with those reported in the psychological literature. Rather than supporting strong conclusions about underlying cognitive mechanisms, these findings highlight the value of perspective transformation as a diagnostic setting for comparing human and model behaviour. At present, our analysis is limited to behavioural outcomes, namely performance across different perspective conditions. Future work could therefore move beyond accuracy-based comparisons and investigate intermediate reasoning traces, such as chain-of-thought explanations, reasoning trajectories, or other forms of model-generated rationales. Such analyses may help clarify whether similarities and differences in task performance reflect comparable representational strategies or fundamentally different computational processes.

Finally, our results may have implications for GeoAI and language-based spatial reasoning systems. Practical applications such as spatial question answering, navigation assistance, and wayfinding support often require information to be translated between route-based and survey-based descriptions. The reduced performance observed in cross-perspective tasks suggests that such transformations remain a challenge even when basic spatial reasoning is largely mastered. This observation motivates future work on architectures that explicitly represent perspective information, as well as hybrid approaches that combine language models with established qualitative spatial reasoning techniques.

8 Conclusion

We begin with three basic spatial reasoning tasks that can be viewed as abstract versions of everyday inferential tasks: the **1D**-, **2D**- and **OMR**-tasks. All language models we investigated achieved satisfactory performance on the **1D**- and **2D**-task, indicating that they can learn to solve these problems. Notably, DeBERTa-v3 copes well with the computationally demanding **OMR**-task. It also demonstrates a certain degree of both in-domain and out-of-domain generalisation. Concentrating on DeBERTa-v3, given its superior performance, we then turn our attention to the issue of perspective transformation, which has long been studied in cognitive psychology on humans. We abstract the perspective transformation task into two tasks: the **2D2OMR**- and **OMR22D**-tasks, switching from survey view (2D) to route view (OMR), or from route view to survey view, respectively. We find that the model is better at switching from route view to survey view, in contrast to the phenomenon observed on humans in psychological experiments. These findings suggest that language models and humans may exhibit different patterns of performance in perspective-dependent spatial reasoning, although the differences in experimental setup preclude strong conclusions regarding the underlying cognitive processes.

References

1 Marco Aiello, Ian Pratt-Hartmann, and Johan Van Benthem, editors. *Handbook of spatial logics*. Dordrecht: Springer, 2007. doi:10.1007/978-1-4020-5587-4.

2 Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.ijcnlp-main.45.

3 Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1075.

4 Ruth M. J. Byrne and P. N. Johnson-Laird. Spatial Reasoning. *Journal of Memory and language*, 28:564–575, 1989. doi:10.1016/0749-596X(89)90013-2.

5 Anthony G. Cohn. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8, 2023. doi:10.48550/arXiv.2309.15577.

6 Anthony G Cohn and Robert E Blackwell. Evaluating the ability of large language models to reason about cardinal directions. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, volume 315 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:9, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.COSIT.2024.28.

7 Anthony G. Cohn and Robert E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions, revisited. In *QR 2025: 38th International Workshop on Qualitative Reasoning at IJCAI*, Montreal, Canada, 2025. doi:10.48550/arXiv.2507.12059.

8 Anthony G Cohn and Jose Hernandez-Orallo. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs, 2023. doi:10.48550/arXiv.2304.11164.

9 Anthony G. Cohn and Jochen Renz. Chapter 13 Qualitative spatial representation and reasoning. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 551–596. Elsevier, 2008. doi:10.1016/S1574-6526(07)03013-1.

10 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

11 Heshan Du, Natasha Alechina, and Anthony G. Cohn. A logic of directions. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1695–1702, Yokohama, Japan, 2020. International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2020/235.

12 Andrew U. Frank. Qualitative spatial reasoning: cardinal directions as an example. *International journal of geographical information systems*, 10(3):269–290, 1996. doi:10.1080/02693799608902079.

13 Christian Freksa. Using orientation information for qualitative spatial reasoning. In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space: International Conference GIS—From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning*, pages 162–178, Pisa, Italy, 1992. Springer Berlin Heidelberg. doi:10.1007/3-540-55966-3_10.

14 Nir Fulman, Abdulkadir Memduhoglu, and Alexander Zipf. Evidence for systematic bias in the spatial memory of large language models. In *GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024*, pages 57–62, Glasgow, Scotland, 2024. Springer Cham. URL: https://ceur-ws.org/Vol-3683/paper8.pdf.

15 Google. Gemini 3.5 flash, 2025. URL: https://deepmind.google/models/model-cards/gemini-3-5-flash/.

16 Gracjan Góral, Alicja Ziarko, Piotr Miłoś, Michał Nauman, Maciej Wołczyk, and Michał Kosiński. Beyond recognition: Evaluating visual perspective taking in vision language models, 2025. doi:10.48550/arXiv.2505.03821.

17 Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou,

Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.1229.

18 Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

19 Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

20 Majid Hojati and Rob Feick. Large language models: Testing their capabilities to understand and explain spatial concepts. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 31:1–31:9, Québec City, Canada, 2024. doi:10.4230/LIPIcs.COSIT.2024.31.

21 Roman Kontchakov, Ian Pratt-Hartmann, and Michael Zakharyaschev. Spatial reasoning with RCC8 and connectedness constraints in Euclidean spaces. *Artificial Intelligence*, 217:43–75, 2014. doi:10.1016/j.artint.2014.07.012.

22 Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

23 Bridget Leonard and Scott O. Murray. Cognitively-inspired tokens overcome egocentric bias in multimodal models, 2026. doi:10.48550/arXiv.2601.16378.

24 Fangjun Li. *Benchmarking and enhancing spatial reasoning in large language models*. PhD thesis, University of Leeds, 2025.

25 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18500–18507, Vancouver, Canada, 2024. AAAI Press, Washington, DC, USA. doi:10.1609/aaai.v38i17.29811.

26 Zekai Li, Amin Farjudian, and Heshan Du. A logic of east and west for intervals. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 17:1–17:8, Québec City, Canada, 2024. doi:10.4230/LIPIcs.COSIT.2024.17.

27 G E Ligozat. Reasoning about cardinal directions. *Journal of Visual Languages and Computing*, 9:23–44, 1998. doi:10.1006/jvlc.1997.9999.

28 Gérard Ligozat, editor. *Qualitative Spatial and Temporal Reasoning*. John Wiley & Sons, 2013. doi:10.1002/9781118601457.

29 Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods, 2025. doi:10.48550/arXiv.2511.15722.

30 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

31 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

32 Tharindu Madusanka, Riza Theresa Batista-Navarro, and Ian Pratt-Hartmann. Identifying the limits of transformers when performing model-checking with natural language. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3539–3550, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.eacl-main.257.

33 Lachlan McPheat, Navdeep Kaur, Robert Blackwell, Alessandra Russo, Anthony G. Cohn, and Pranava Madhyastha. DecompSR: A dataset for decomposed analyses of compositional multihop spatial reasoning, 2026. doi:10.48550/arXiv.2511.02627.

34 Roshanak Mirzaee and Parisa Kordjamshidi. Transfer learning with synthetic corpora for spatial role labeling and reasoning, 2022. doi:10.48550/arXiv.2210.16952.

35 Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.364.

36 Reinhard Moratz and Marco Ragni. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1):75–98, 2008. Spatial and Image-based Information Systems. doi:10.1016/j.jvlc.2006.11.001.

37 Reinhard Moratz, Jochen Renz, and Diedrich Wolter. Qualitative spatial reasoning about line segments. In *European Conference on Artificial Intelligence*, 2000. URL: https://api.semanticscholar.org/CorpusID:1424882.

38 Till Mossakowski and Reinhard Moratz. Qualitative reasoning about relative direction of oriented points. *Artificial Intelligence*, 180-181:34–45, 2012. doi:10.1016/j.artint.2011.10.003.

39 Till Mossakowski and Reinhard Moratz. Relations between spatial calculi about directions and orientations. *Journal of Artificial Intelligence Research*, 54(1):277–308, September 2015. doi:10.1613/jair.4631.

40 Joseph O'Rourke, editor. *Computational geometry in C*. Cambridge university press, 2012. doi:10.1017/CBO9780511804120.

41 Tanawan Premsri and Parisa Kordjamshidi. Tuning language models with spatial logic for complex reasoning. In *The 4th International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, 2024. URL: https://openreview.net/forum?id=cDm7zGenhZ.

42 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21:1–67, 2020. URL: http://jmlr.org/papers/v21/20-074.html.

43 Marco Ragni and Markus Knauff. A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, 120:561–588, 2013. doi:10.1037/a0032460.

44 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi:10.18653/v1/D16-1264.

45 David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 165–176, Cambridge, United States, 1992. Morgan Kaufmann Publishers Inc.

46 Kyle Richardson and Ashish Sabharwal. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11209–11219, Virtual, 2022. AAAI Press, Palo Alto, California USA. doi:10.1609/aaai.v36i10.21371.

47 Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Seoul, South Korea, 2025. doi:10.48550/arXiv.2507.07644.

48 Felipe Salvatore, Marcelo Finger, and Roberto Hirata Jr. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for*

Low-Resource NLP (DeepLo 2019), pages 22–30, Stroudsburg, PA, USA, May 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-6103.

49 Amy L. Shelton and Timothy P. McNamara. Orientation and perspective dependence in route and survey learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30:158–170, 2004. doi:10.1037/0278-7393.30.1.158.

50 Amy L. Shelton and Holly A. Pippitt. Fixed versus dynamic orientations in environmental learning from ground-level and aerial perspectives. *Psychological Research*, 71:333–346, 2007. doi:10.1007/s00426-006-0088-9.

51 Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11321–11329, Virtual, 2022. AAAI Press, Palo Alto, California USA. doi:10.1609/aaai.v36i10.21383.

52 Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.1195.

53 Alexander W. Siegel and Sheldon H. White. The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, 10:9–55, 1975. doi:10.1016/S0065-2407(08)60007-5.

54 Michael Sioutis and Diedrich Wolter. Qualitative spatial and temporal reasoning: Current status and future challenges. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21): Survey Track*, 2021. doi:10.24963/ijcai.2021/624.

55 Holly A. Taylor and Barbara Tversky. Spatial mental models derived from survey and route descriptions. *Journal of Memory and language*, 31:261–292, 1992. doi:10.1016/0749-596X(92)90014-O.

56 Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.303.

57 Barbara Tversky. Spatial Mental Models. *Psychology of learning and motivation*, 27:109–145, 1991. doi:10.1016/S0079-7421(08)60122-X.

58 Barbara Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *International Conference on Spatial Information Theory (COSIT 1993)*, pages 14–24, Marciana Marina, Elba Island, Italy, 1993. Springer Berlin, Heidelberg. doi:10.1007/3-540-57207-4_2.

59 Maijunxian Wang, Yijiang Li, Bingyang Wang, Tianwei Zhao, Ran Ji, Qingying Gao, Emmy Liu, Hokin Deng, and Dezhi Luo. Egocentric bias in vision-language models, 2026. doi:10.48550/arXiv.2602.15892.

60 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

61 Steffen Werner, Bernd Krieg-Brückner, Hanspeter A. Mallot, Karin Schweizer, and Christian Freksa. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. In *Informatik '97 Informatik als Innovationsmotor*, pages 41–50, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. doi:10.1007/978-3-642-60831-5_8.

62 Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks, 2015. arXiv:1502.05698.

63 Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1101.

64 Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6:1–30, 2025. doi:10.1016/j.patter.2025.101370.

65 Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. Evaluating large language models on spatial tasks: A multi-task benchmarking study, 2025. doi:10.48550/arXiv.2408.14438.

66 Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition, 2025. doi:10.48550/arXiv.2511.21471.

67 Anran Yang, Cheng Fu, Qingren Jia, Weihua Dong, Mengyu Ma, Hao Chen, Fei Yang, and Hui Wu. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, 39(9):2009–2044, 2025. doi:10.1080/13658816.2025.2490701.

68 Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, 2019. Curran Associates, Inc. doi:10.48550/arXiv.1906.08237.

69 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=5Xc1ecxO1h.

A Zero-shot Prompt

Input: *the_original_input*

Question: Each uppercase letter represents a point, and different points cannot overlap. Can all statements in the input be true at the same time? Only answer Yes or No.

Answer:

LMM Modality Transfer: A Pre-Requisite for Autonomous GIS Agents

Ivan Majic[1]
Graz University of Technology, Austria
University of Vienna, Austria

Zexian Huang
Graz University of Technology, Austria

Franziska Hübl
Graz University of Technology, Austria

Krzysztof Janowicz
University of Vienna, Austria

Meilin Shi
University of Liverpool, UK

Mina Karimi
University of Vienna, Austria

Zilong Liu
University of Vienna, Austria

Alexandra Fortacz-Lazan
University of Vienna, Austria

Abstract

AI models are becoming increasingly adept at understanding and processing spatial information, thereby facilitating agentic problem-solving in spatial tasks and workflows. However, most of the research on their spatial capabilities (e.g., spatial reasoning) has focused on the textual modality as input and output. This contrasts with the human approach to GIS workflows, where text and visual modalities are often used together, interchangeably, and in a complementary manner. Thus, to truly achieve an automated GIS analysis pipeline or carry out human-designed GIS workflows, AI models – Large Multimodal Models (LMMs) in particular – need to be able to seamlessly transition between image- and text-based modalities that are traditionally used in such workflows. We present a modality transfer task that (1) asks an LMM to first describe an input image of colored squares in a regular grid, and (2) asks a new LMM instance to re-generate an image of the original spatial scene using the textual description output by the former model. This task quantifies the ability of LMMs to transfer spatial information between image and text modalities. Ultimately, by examining the modality transfer capability of LMMs through the lens of spatial information theory, this work highlights a critical bottleneck: achieving strong and robust geospatial understanding in LMMs requires rigorous, multi-modal alignment. Our results indicate that recent LMMs (here from OpenAI) still struggle with modality transfer, when tasked with re-generating an image of a simple spatial grid of color squares.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Intelligent agents; Information systems → Geographic information systems

Keywords and phrases Large Multimodal Model (LMM), Spatial Reasoning, GIS Agent, Modality Transfer, GeoAI

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.7

Supplementary Material *Software (Source Code)*:
`https://github.com/Geoinfo-TUGraz/COSIT2026_LMM_modality_transfer`
archived at `swh:1:dir:4d079d748cdb755f77f630344e664c8a8c9e65f7`

1 Introduction

Typical GIS workflows inherently use both text and image modalities interchangeably, depending on the situation and the communication needs. For example, two human GIS analysts who work side by side on a joint project may exchange textual information like

[1] corresponding author

© Ivan Majic, Zexian Huang, Franziska Hübl, Krzysztof Janowicz, Meilin Shi, Mina Karimi, Zilong Liu, and Alexandra Fortacz-Lazan;
licensed under Creative Commons License CC-BY 4.0

feature attributes, but also discuss the semantic meaning of certain colored polygons by asking questions like *"what does this elongated purple shape represent"*. If one or both of these GIS analysts were to be replaced by autonomous GIS agents, we argue that the same (human-expert) level of understanding and seamless transition between both modalities of geographic information is a pre-requisite for smooth, reliable, and deterministically repeatable operation [19].

Development of such intelligent GIS agents is a continuation of the long-standing desire and need to automate GIS workflows. Tasks such as information retrieval, data preparation, and data cleaning have always been resource-hungry. This is why many computational workflows and task-specific GIS automations have been developed as plugins to GIS environments such as ArcGIS or QGIS [15], or as standalone software tools [13, 14]. With the advent of generative AI and foundation models that are capable of addressing a variety of downstream tasks without specific pre-training, the aspirations of autonomous GIS agents have also risen [10]. An intelligent GIS agent that can understand the data it is working with and autonomously adjust the workflows to the specific situation would be of great value. Hence, such a *GeoMachina* has been declared one of the potential common moonshots for the field of GeoAI [7]. While there is substantial progress [10, 1], some fundamental considerations regarding the multi-modality of spatial information remain unaddressed.

Thus, we bring forth an essential question regarding the underlying foundation model infrastructure and its readiness to fully autonomously handle inherently multi-modal GIS workflows: *"how well can large multimodal models (LMMs) understand each other and can they transfer spatial information between the two modalities (image and text) without losing information and quality?"* To this end, we have designed a modality transfer task that takes images of (spatially arranged) grids of colored cells (the likes of which can be found in typical land cover maps), prompts the LMM to textually describe the spatial configuration visible in the image, and then asks another model to generate an image based on this textual output. In other words, the proposed task assesses the ability of LMMs to perform `image` → `text` → `image` modality transfer. Our case study applies this task on recent-generation OpenAI models with different grid sizes and numbers of colors. **Thus, the contributions of this study are two-fold**:

- We identify the lossless modality transfer as a fundamental requirement for many GIS tasks to be solved by autonomous GIS agents.
- We define a modality transfer task that can be used to benchmark the progress of spatial capabilities of LMMs in this context.

Thereby, our work addresses the important challenge of aligning the progress of the state-of-the-art foundational AI models with the context of GIScience by testing their ability to comprehend geographical information.

The remainder of this paper is organized as follows. In Section 2, we review the literature related to our study. We propose our approach in Section 3. Then, the results are explained in Section 4. We discuss our finding in Section 5. Finally, Section 6 concludes the paper.

2 Related work

The rapid evolution of GeoAI has significantly advanced geographic knowledge discovery and spatial analysis. Previous foundational work has highlighted the transformative potential of spatially explicit AI techniques in addressing complex geographic problems, moving beyond traditional analysis to intelligent, agentic workflows [7, 12]. While they outline various fundamental challenges in GeoAI, in particular, spatial heterogeneity, scale dependencies, and

the need for spatially explicit models, the specific challenges of *large multi-modal* geographic reasoning remain largely unaddressed. In contrast, single-mode spatial reasoning has received increasing attention [5, 8, 4, 6]. Work has also focused on vision language models and their ability to answer spatial reasoning questions [2]. Recently, Xie et al. demonstrated that LLMs still lack a deeper understanding of spatial relations but rely on linguistic patterns instead [20].

Traditional geographic information and human-driven GIS workflows are inherently multi-modal. Human analysts seamlessly blend visual information (e.g., maps, satellite imagery, and topological configurations) with textual information (e.g., attributes, spatial queries, and descriptive metadata). However, even recent GeoAI and foundation model research evaluates these modalities independently [8]. Recent work has also shown that foundation models yield uneven success depending on which of these modalities the spatial reasoning task is presented in [16], showcasing there is a difference in spatial reasoning capabilities in different modalities. **Our work directly addresses this gap.** By introducing a framework to explicitly test the multi-modal spatial reasoning capabilities of LMMs, we study the lossless transfer of spatial information between visual and textual modalities as a critical, unaddressed prerequisite for the development of fully autonomous GeoAI agents.

Evaluating the true representation and reasoning capabilities of AI systems requires benchmarks that isolate genuine problem-solving from mere pattern matching or data memorization. A prominent example is the Abstract and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) introduced by Chollet et al. [3]. A guiding design philosophy of ARC-AGI is to formulate tasks that are intuitive for humans yet highly challenging for AI models, thereby exposing gaps and weaknesses in an AI's capacity to generalize abstract rules from limited examples without prior training. Our proposed modality transfer task shares this core principle but targets a different aspect. While ARC-AGI focuses on abstract logical reasoning and skill acquisition within a single visual-symbolic domain, our modality transfer task explicitly evaluates the *transferability and preservation of spatial information across disparate data modalities*.

Human spatial cognition can effortlessly translate visual cues into language descriptions and vice versa. For instance, one person can look at a visual grid and verbally describe a "*red square situated in the top-left corner*", and another one can accurately reproduce those visual cues solely based on the description. Our `ground-truth image` → `generated text` → `generated image` pipeline isolates this exact translation. By abstracting the task away from heavily prior GIS knowledge (such as specific coordinate reference systems or complex vector geometries) and utilizing colored spatial grids, we ensure a fair assessment. Similar to ARC-AGI, our task relies on basic spatial core priors (grids, colors, and relative positioning). However, instead of asking AI models to deduce a hidden logical rule, we test LMM's capability to accurately translate spatial reality from one modality to another.

3 Approach

In order to assess the ability of the LMM to understand spatial information, we propose an approach that transfers the modality of a given spatial land-cover-like example from an image representation – into a textual representation – and back into an image representation (i.e., `image` → `generated text` → `image` modality transfer). We call this the *modality transfer task*. By performing it, we can get an estimation of how well the model really understands the provided spatial configuration. We can also directly compare the final output (generated image) to the original input (image), as they are both in the same modality and, when the system performs completely correctly, should be identical.

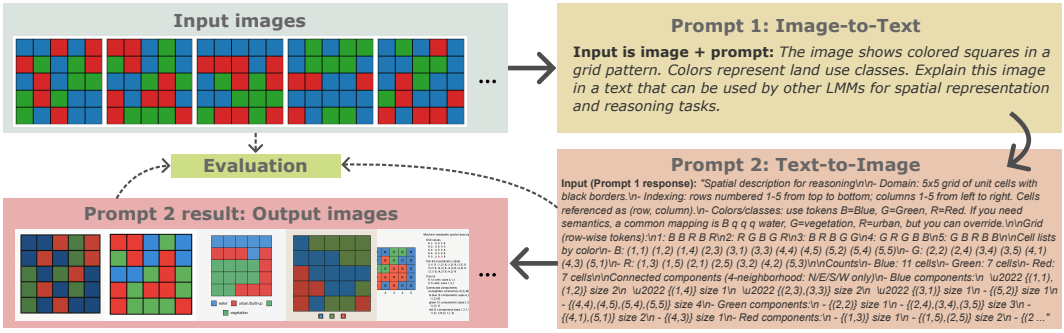


Figure 1 A diagram showing the flow of the proposed modality transfer task. Input images are generated in different configurations and batch sizes. They are then used as an input to the Prompt 1, along with a textual prompt that directs the AI how to generate a suitable textual description. The outputs of the Prompt 1 are used as inputs for the Prompt 2 where textual descriptions are used to generate new images. Finally, the generated outputs of both Prompt 1 (text) and Prompt 2 (images) are evaluated against the original input images.

The overall approach to the LMM modality transfer task is carried out as a multi-stage workflow designed to isolate, control, and quantify each step of the task. As shown in Figure 1, the workflow consists of the generation of input ground-truth images (i.e., $N \times N$ grids of colored rectangles), an LMM prompting sequence (image-to-text followed by text-to-image), the systematic extraction of generated images into discrete matrices, and a quantitative evaluation utilizing perceptual color thresholding of color matrices, text-based and image-based Levenshtein Distance [9] and, Earth Mover's Distance (EMD) [18].

While Figure 1 shows the logical flow of the workflow, the approach is implemented in Python programming language as a series of interactive Python notebooks. The repository containing the code, input images, and LMM prompt outputs (text and images) is published in a freely accessible repository[1].

3.1 Generation of input (baseline) images

The ground-truth images consisting of $N \times N$ grids of colored rectangles are generated procedurally based on the Tableau color palette available in the Matplotlib Python library. The five selected colors from this 10-color palette are depicted in Figure 2, while the perceptual differences between them are expressed using the CIEDE2000 (ΔE_{00}) [11] color difference formula which can express the perceived difference between two colors better than any measure based on the RGB or HSL color values [17]. These distinct colors can serve as semantic labels for discrete categorical GIS data, such as land use classes, ensuring that different categories are visually distinguishable and minimizing perceptual ambiguity.

To evaluate varying degrees of spatial complexity, the grid size N is incrementally scaled from 5 to 10, and the number of colors used was increased from 3 to 4 and 5 (i.e., red, green, blue, orange, and cyan), where each cell was randomly assigned one of the possible color values (Figure 3). Both the images and their corresponding exact spatial matrices are recorded to serve as the definitive baseline for all subsequent comparisons.

[1] `https://github.com/Geoinfo-TUGraz/COSIT2026_LMM_modality_transfer.git`

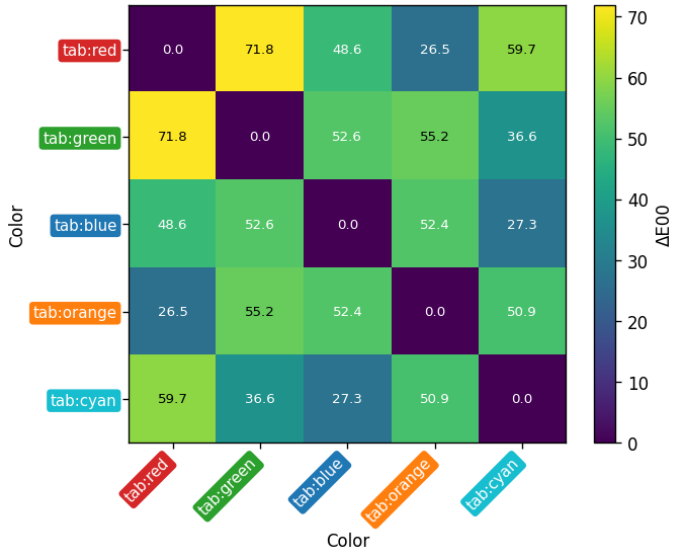


Figure 2 The five selected named colors of the Tableau Palette available in the Python Matplotlib library and the differences between them expressed as ΔE_{00}.

Listing 1 Prompt 1: Image-to-Text.

```
The image shows colored squares in a
grid pattern. Colors represent land use classes. Explain this image
in a text that can be used by other LMMs for spatial representation
and reasoning tasks.
```

3.2 Prompting LMMs

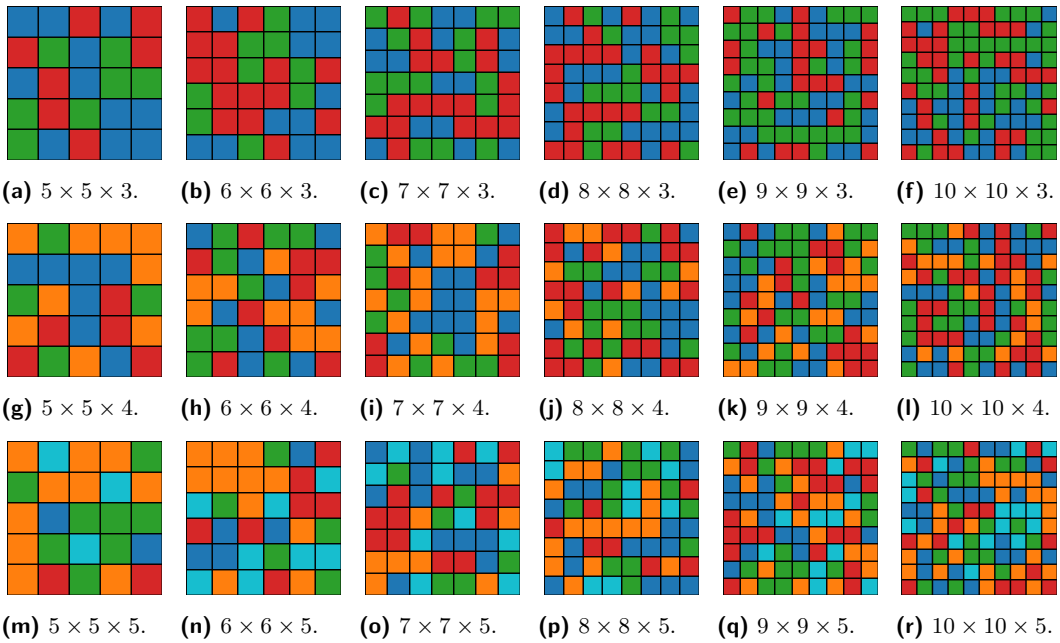
3.2.1 Prompt 1: Image-to-Text

As depicted in Figure 1, generated input images are provided to an LMM. The accompanying prompt directs the LMM to act as a spatial extraction agent, tasking it to generate a machine-readable textual representation of the spatial layout that details the domain size, indexing system, and the coordinate mapping of the colored cells. The exact prompt was written as follows:

As a response, the LMM would return the generated textual description of the input image. Figure 4 shows an example of the prompt response when image shown in Figure 3a was used as an input.

3.2.2 Prompt 2: Text-to-Image

In the subsequent step, the textual description output by Prompt 1 is used as an input for Prompt 2 where an image-generation variant of an LMM is prompted. No additional system prompts or directions were used, as the input textual description is usually already quite detailed and the image-generation LMM understands that by default its task is to generate an image output based on the textual input. Figure 5 shows the generated output images that correspond to the input images from Figure 3.



(a) $5 \times 5 \times 3$. **(b)** $6 \times 6 \times 3$. **(c)** $7 \times 7 \times 3$. **(d)** $8 \times 8 \times 3$. **(e)** $9 \times 9 \times 3$. **(f)** $10 \times 10 \times 3$.

(g) $5 \times 5 \times 4$. **(h)** $6 \times 6 \times 4$. **(i)** $7 \times 7 \times 4$. **(j)** $8 \times 8 \times 4$. **(k)** $9 \times 9 \times 4$. **(l)** $10 \times 10 \times 4$.

(m) $5 \times 5 \times 5$. **(n)** $6 \times 6 \times 5$. **(o)** $7 \times 7 \times 5$. **(p)** $8 \times 8 \times 5$. **(q)** $9 \times 9 \times 5$. **(r)** $10 \times 10 \times 5$.

Figure 3 Examples of baseline images generated for this study. The grid sizes range from N=5 on the left to N=10 on the right, and the number of colors randomly assigned to each cell range from N_colors=3 in the first row to N_colors=5 in the third row.

3.3 Processing of LMM outputs

To be able to compare the outputs of Prompt 1 and Prompt 2 to the input images, some preprocessing steps are required to sample and extract the relevant information and convert it into the format that allows the evaluation measures to be calculated. For this purpose, we converted all three color grid depictions – input images, generated texts, and generated images – into 2-dimensional arrays of color names. This required separate processing steps for the textual and image outputs of the LMM prompts.

The textual description outputs of Prompt 1 (image-to-text) usually contained longer natural language descriptions of the grids visible in the input images that explained the size, format, colors used, color meanings, clusters of the same colors, and sometimes even topological relations between individual clusters. However, the majority of outputs also included some kind of matrix encoding of the input grid and this is the information that we have extracted and standardized into 2-dimensional arrays of colors where each color is represented by a single letter.

The outputs of Prompt 2 (text-to-image) are in an image modality and require a different kind of processing to convert the information into a 2D array. Although it is possible to perform this step with yet another LMM prompt, this would be subject to the same uncertainty as Prompt 1, as it would require another image-to-text prompt and would not seem reasonable in a workflow that is intended to evaluate LMMs in the first place. Thus, we have created a small application that allows user to sample colors from the cells of output images using a graphical user interface (Figure 6). It should be noted that output images can also contained highly distorted grids that cannot be sampled and have to be omitted completely (see Table 1). After all output images from one directory have been sampled, the application converts the sampled points and colors into 2D arrays.

```
Spatial description for reasoning
- Domain: 5×5 grid of unit cells with black borders.
- Indexing: rows numbered 1→5 from top to bottom; columns 1→5 from left to right. Cells referenced as (row, column).
- Colors/classes: use tokens B=Blue, G=Green, R=Red. If you need semantics, a common mapping is B→water, G→vegetation,
R→urban, but you can override.
Grid (row-wise tokens):
1: B B R B R
2: R G B G R
3: B R B G G
4: G R G B B
5: G B R B B
Cell lists by color
- B: (1,1) (1,2) (1,4) (2,3) (3,1) (3,3) (4,4) (4,5) (5,2) (5,4) (5,5)
- G: (2,2) (2,4) (3,4) (3,5) (4,1) (4,3) (5,1)
- R: (1,3) (1,5) (2,1) (2,5) (3,2) (4,2) (5,3)
Counts
- Blue: 11 cells
- Green: 7 cells
- Red: 7 cells
Connected components (4-neighborhood: N/E/S/W only)
- Blue components:
...
- All three color pairs have edge adjacencies somewhere in the grid (B–G, B–R, G–R).
- The largest contiguous patch is Blue of size 4 at the bottom-right corner: {(4,4),(4,5),(5,4),(5,5)}.
This text is self-contained and can be parsed to reconstruct the grid and reason about spatial relations, coverage, and
connectivity.
```

Figure 4 Textual description of the image shown in Figure 3a, generated with the OpenAI's GPT5 model.

3.4 Evaluation of LMM spatial hallucination

Evaluating the accuracy of the modality transfer requires quantifying both spatial and attribute degradation. Generative image models rarely output exact, deterministic hex color codes, often producing slight gradients or perceptual variations. Facing this challenge, we apply a strict CIEDE2000 (ΔE_{00}) thresholding function to rigorously evaluate the extracted visual matrices against the discrete ground-truth grids. Let $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$ denote the ground-truth color palette. For each generated pixel color c, this heuristic calculates the perceptual distance $\Delta E_{00}(c, p_i)$ to every palette color. Generated colors falling within a defined perceptual threshold τ are mapped to their nearest ground-truth categorical class, while those exceeding the threshold are penalized and registered as out-of-bounds attribute hallucinations (h). The assigned class l for a generated color c is defined as:

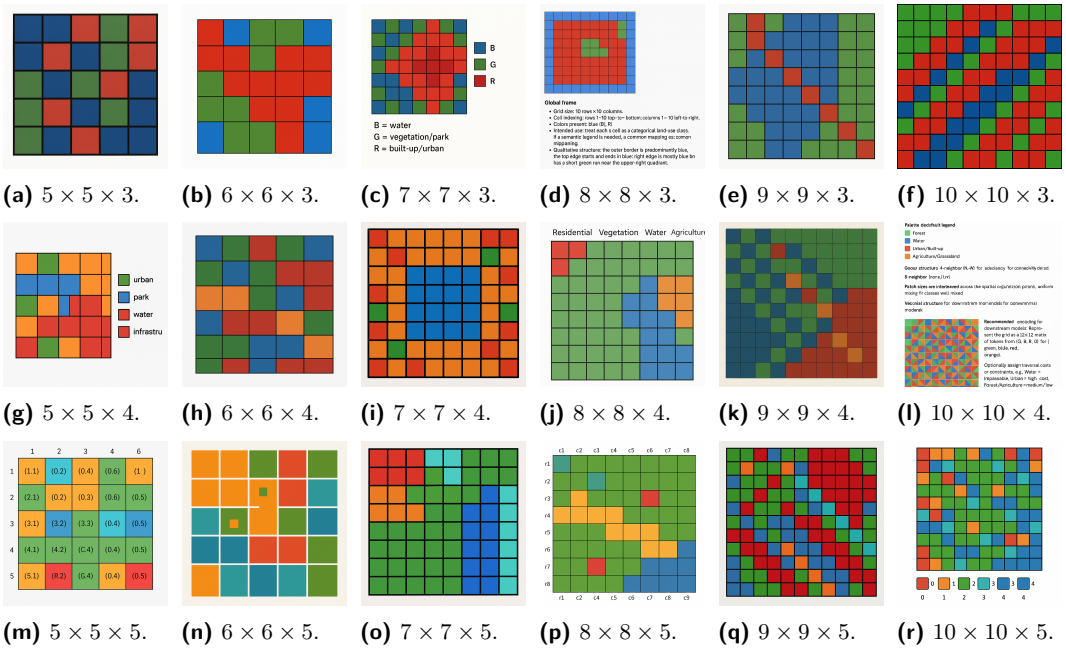
$$l = \begin{cases} \arg\min_{p_i \in \mathcal{P}} \Delta E_{00}(c, p_i) & \text{if } \min \Delta E_{00}(c, p_i) \leq \tau \\ h & \text{otherwise} \end{cases}$$

3.4.1 1D Levenshtein Distance

Following their categorical normalization into discrete matrices, let M represent the ground-truth matrix and $\hat{M}$ represent the generated matrix. To facilitate the evaluation of spatial determinism, these 2D spatial structures are flattened into 1D sequential strings, S and $\hat{S}$, such that $S = (s_1, s_2, \ldots, s_{N^2})$. Subsequently, the Levenshtein Distance [9] is computed between the generated sequences and the ground truth to strictly quantify spatial loss. Let $\mathcal{L}_{S,\hat{S}}(|S|, |\hat{S}|)$ denote the distance between the two sequences, defined recursively to calculate the minimum number of mathematical edit operations required to reconstruct the original input sequence:

$$\mathcal{L}_{S,\hat{S}}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathcal{L}_{S,\hat{S}}(i-1,j) + 1 \\ \mathcal{L}_{S,\hat{S}}(i,j-1) + 1 \\ \mathcal{L}_{S,\hat{S}}(i-1,j-1) + \mathbb{I}(s_i \neq \hat{s}_j), \end{cases} & \text{otherwise} \end{cases}$$

where $\mathbb{I}$ is the indicator function that equals 1 when the characters differ and 0 when they match. Within this spatial context, the operations directly translate to model errors: insertions ($\mathcal{L}_{S,\hat{S}}(i,j-1)+1$) represent hallucinated blocks, deletions ($\mathcal{L}_{S,\hat{S}}(i-1,j)+1$)



(a) $5 \times 5 \times 3$. **(b)** $6 \times 6 \times 3$. **(c)** $7 \times 7 \times 3$. **(d)** $8 \times 8 \times 3$. **(e)** $9 \times 9 \times 3$. **(f)** $10 \times 10 \times 3$.

(g) $5 \times 5 \times 4$. **(h)** $6 \times 6 \times 4$. **(i)** $7 \times 7 \times 4$. **(j)** $8 \times 8 \times 4$. **(k)** $9 \times 9 \times 4$. **(l)** $10 \times 10 \times 4$.

(m) $5 \times 5 \times 5$. **(n)** $6 \times 6 \times 5$. **(o)** $7 \times 7 \times 5$. **(p)** $8 \times 8 \times 5$. **(q)** $9 \times 9 \times 5$. **(r)** $10 \times 10 \times 5$.

Figure 5 Examples of LMM generated output images based on input baseline images from Figure 3 after they have been through the image-to-text transfer with Prompt 1 and subsequently through the text-to-image transfer with Prompt 2.

represent missing blocks, and substitutions ($\mathbb{I}(s_i \neq \hat{s}_j) = 1$) represent incorrectly placed or colored blocks. If the input and output images are identical (given the allowed perceptual difference threshold for colors τ), the Levenshtein Distance between them is zero, i.e., $\mathcal{L}_{S,\hat{S}}(|S|, |\hat{S}|) = 0$.

3.4.2 2D Spatial Earth Mover's Distance

While the 1D Levenshtein sequence evaluation effectively captures topological errors of generated images and texts, this evaluation is inherently insensitive to the geometric magnitude of spatial displacements. To quantify the true 2D geographic loss and penalize severe spatial hallucinations, we formulate a discrete, class-wise Spatial Earth Mover's Distance (EMD) [18] as an optimal transport problem.

Let $\mathcal{U}$ denote the union of all categorical labels present in both the ground-truth matrix M and the generated matrix $\hat{M}$. For each label $l \in \mathcal{U}$, we extract the sets of 2D Cartesian coordinates representing the spatial locations of those specific class blocks: $C_l = \{(x, y) \mid M_{x,y} = l\}$ and $\hat{C}_l = \{(x, y) \mid \hat{M}_{x,y} = l\}$. Let $n_l = |C_l|$ and $\hat{n}_l = |\hat{C}_l|$ denote the cardinality of these sets.

To account for unmapped features, in particular deletions where the LMM fails to generate a required block, or insertions where the model hallucinates a new, out-of-bounds color, a maximum spatial penalty $\rho = \sqrt{H^2 + W^2}$ is defined, which represents the maximum diagonal distance of the $H \times W$ spatial grid, where H and W stand for height and width of the grid.

In particular, for a given class l, the spatial transport cost $W(C_l, \hat{C}_l)$ computes the minimum Euclidean distance required to align the generated blocks with the ground truth, penalized by any cardinality mismatch:

$$W(C_l, \hat{C}_l) = \min_{\pi} \sum_{i=1}^{\min(n_l, \hat{n}_l)} \|c_i - \hat{c}_{\pi(i)}\|_2 + |n_l - \hat{n}_l|\rho,$$

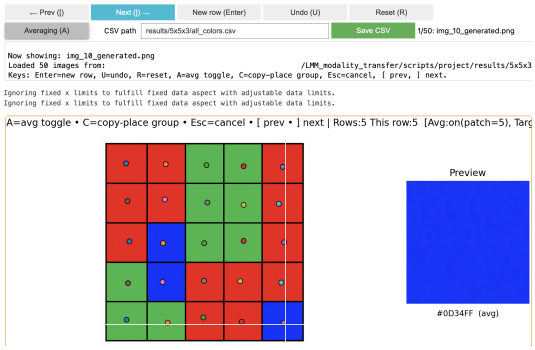


Figure 6 Graphical user interface of the output image color-sampling application. Each point represents a location where the user has sampled the cell color in order to capture the grid layout and colors of an image generated by an LMM.

where π represents the optimal injective mapping from the smaller coordinate set to the larger coordinate set, solved via the Hungarian algorithm (linear sum assignment), and $\|\cdot\|_2$ denotes the standard L_2 Euclidean distance. Finally, the aggregated Spatial EMD is divided by the total number of valid ground-truth blocks, $N_{gt} = \sum_{l \in \mathcal{U}} n_l$, which ensures that the final metric represents the average physical pixel shift per correct block:

$$\text{EMD} = \frac{1}{N_{gt}} \sum_{l \in \mathcal{U}} W(C_l, \hat{C}_l).$$

3.4.3 Interpretation of Metrics

To systematically evaluate the modality transfer pipeline and diagnose the exact nature of spatial degradation, we analyze the interplay between the 1D sequence error (i.e., 1D Levenshtein) and the 2D geometric error (i.e., 2D spatial EMD). While the Levenshtein distance effectively measures whether the generative model understood the topological shape and categorical sequence of the prompt, it is highly sensitive to minor structural shifts and heavily penalizes simple translations. Conversely, the spatial EMD quantifies the absolute geographic magnitude of these displacements, indicating whether the model adheres to strict coordinate geometry.

By evaluating these two metrics, as summarized in Table 1, we establish a diagnostic framework to classify the spatial reasoning failure of LMMs. This decoupled approach allows us to definitively distinguish between minor spatial coordinate drifts, where the model maintains relative topology but fails at absolute positioning and catastrophic generative breakdowns, where all spatial context is completely hallucinated.

4 Results

We performed an experiment where recent generation OpenAI LMMs were evaluated with the modality transfer task proposed in Section 3. We first generated input images with the grid sizes ($N \times N$) ranging from $N = 5$ to $N = 10$, and the number of colors ranging from $N_{colors} = 3$ to $N_{colors} = 5$. For each parameter combination (e.g., $5 \times 5 \times 3$), 50 distinct baseline images were created, resulting in a total of 900 input images. For Prompt 1 (image-

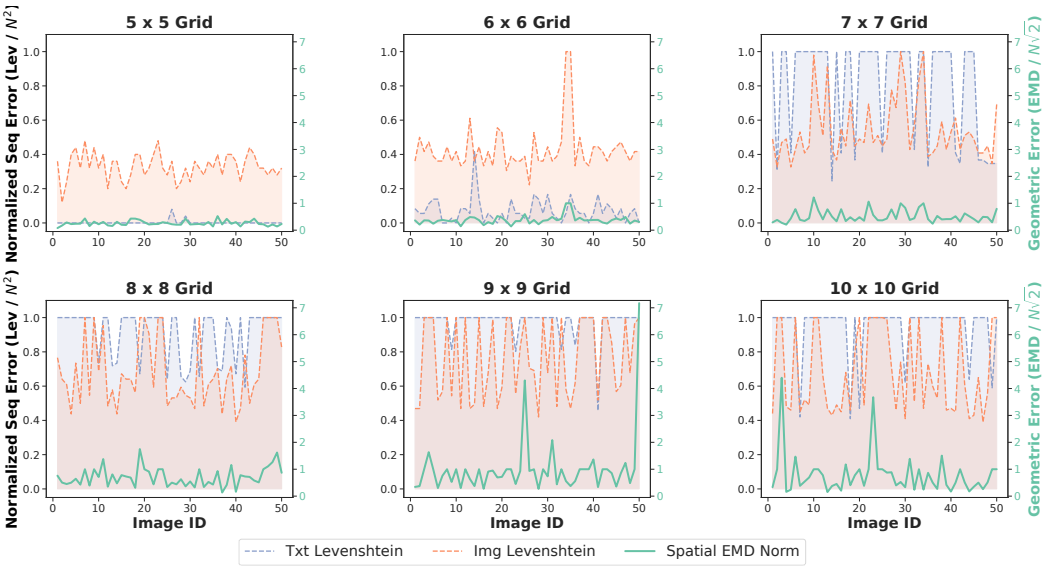
Table 1 Spatial Error Diagnostic Matrix based on Levenshtein Distance and Spatial EMD.

Levenshtein	Spatial EMD	Diagnostic Category & Interpretation
Low	Low	**Minor Attribute Error:** The model successfully preserves sequence and geometry. Represents flawless transfer (if both are exactly 0) or perfect spatial layouts with minor categorical substitutions (e.g., incorrect colors).
High	Low	**Slight Misalignment (Topology Failure):** The model preserves general shape and clustering but lacks absolute coordinate precision. Minor translations (e.g., a 1-cell shift) heavily penalize the 1D sequence while keeping 2D geometric loss low.
High	High	**Severe Hallucination (Catastrophic Failure):** A complete loss of geometric and sequential context. Generated blocks are scattered randomly, omitted entirely, or rendered with out-of-bounds, hallucinated colors.
Low	High	**Geometric Outlier (Shape Distortion):** A mathematically anomaly where the 1D sequence remains mostly intact, but physical boundary violations (e.g., hallucinating a "jagged" grid) incur a massive maximum-diagonal geometric penalty.

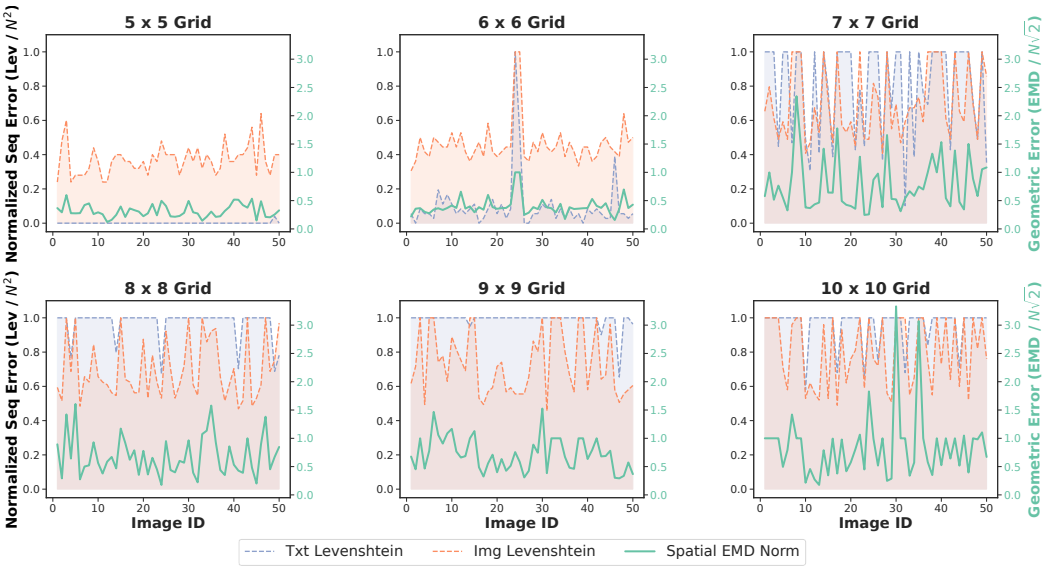
to-text), we utilize OpenAI's `gpt-5` model[2] to generate textual descriptions of input images with $temperature = 1$ and default token size. For Prompt 2 (text-to-image), we use OpenAI's `gpt-image-1` model[3] to generate new images based on the outputs of Prompt 1. We set the parameters for Prompt 2 as $image_size = 1024 \times 1024$ (pixels) and $quality = high$. It is important to note that OpenAI's API is a paid service and image generation prompts may accumulate particularly high costs. We executed prompts for $N_{colors} = 3$ cases by sending API calls directly from our Python scripts. For all other cases, we executed prompts in batches directly on the OpenAI's API dashboard in order to lower the costs and save resources. Total cost of the entire experiment – 900 `gpt-5` prompts, 900 `gpt-image-1` prompts, as well as a number of test prompts – was approximately 200 USD.

Figure 7 shows the evaluation of each individual input image (Image ID axis) with the Normalized Sequential Error (Levenshtein Distance divided by grid size N^2) and the Normalized Geometric Error (Spatial EMD divided by grid diagonal $N\sqrt{2}$) measures. The tested models demonstrate a high degree of variability dependent on both the scale of the grid (N) and the color complexity. In the simplest configuration, testing with 3 color palette (Figure 7a), the vision-to-text sequence error remains relatively low for $N \leq 6$. However, it exhibits a sharp capability drop at $N = 7$, driving downstream errors that peak at $N = 9$. Expanding the palette to 4 colors (Figure 8b) complicates the representational mapping and accelerates this breakdown; baseline vision parsing degrades much earlier, compounding sequence penalties and geometric misalignments. At maximum complexity with 5 colors (Figure 8c), the tested LMMs experience immediate catastrophic failures, where converging high sequence penalties and very high Spatial EMD values indicate a total loss of topological and geometric structure. Across all configurations, scaling spatial density consistently triggers an abrupt loss of geometric determinism.

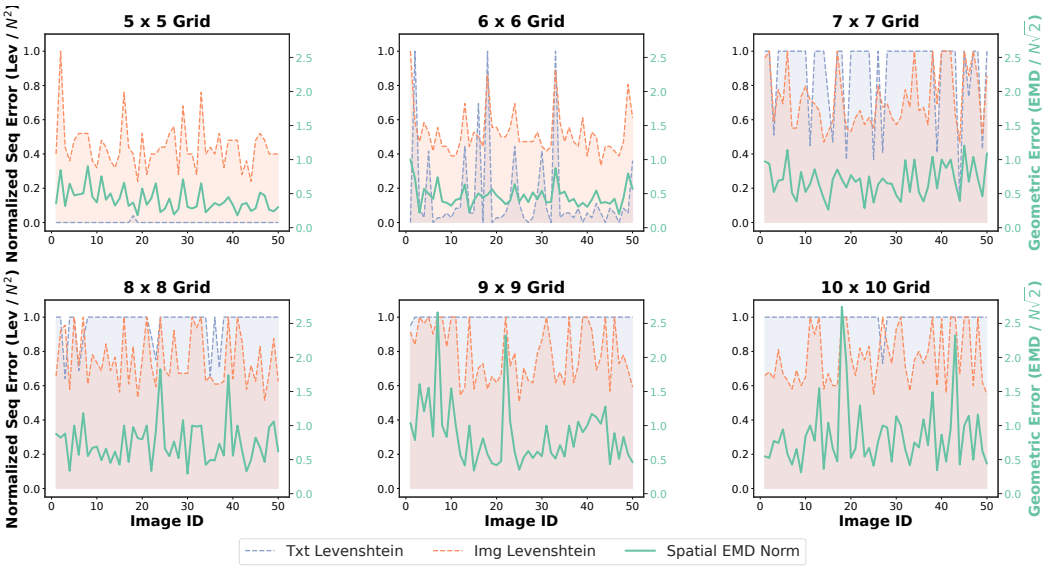
[2] `https://developers.openai.com/api/docs/models/gpt-5`
[3] `https://developers.openai.com/api/docs/models/gpt-image-1`



(a) Evaluation metrics for spatial grids utilizing a ***3-color palette*** (i.e., red, green, blue). At this lower number of grid and color complexity, the models maintain relatively stable topological and geometric coherence at smaller grid sizes ($N \leq 7$). While slight misalignments occur, catastrophic geometric failures (high EMD) remain infrequent until the grid size reaches $N \geq 9$.

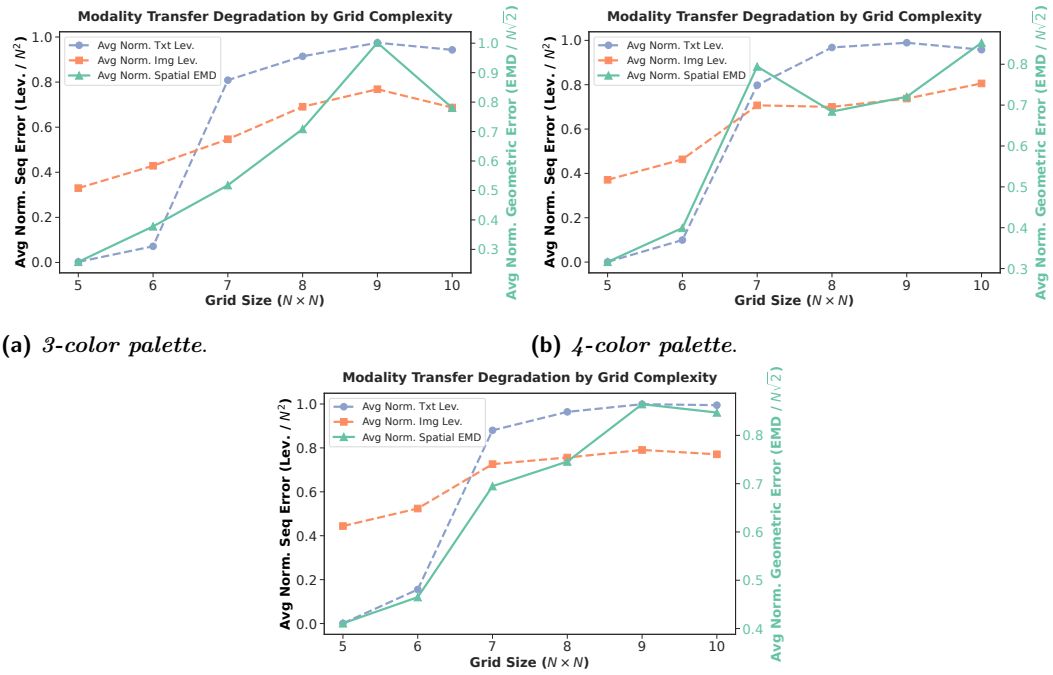


(b) Evaluation metrics for spatial grids utilizing a ***4-color palette*** (i.e., red, green, blue, orange). The introduction of a fourth categorical color significantly destabilizes the modality transfer. Notably, at $N \geq 8$, the models exhibit simultaneous, severe spikes in both sequence error (Img Levenshtein) and geometric displacement (Spatial EMD), indicating complete losses of spatial context within individual samples.



(c) Evaluation metrics for spatial grids utilizing a ***5-color palette*** (i.e., red, green, blue, orange, cyan). Maximum visual complexity results in immediate and significant modality transfer breakdowns across nearly all grid sizes. The baseline vision-to-text parsing (Txt Levenshtein) exhibits frequent failures early on, subsequently corrupting the text-to-image generation phase and driving continuous catastrophic geometric errors.

Figure 7 Detailed evaluation of modality transfer performance across varying grid sizes ($N = 5$ to $N = 10$) and color palette complexities (3, 4, and 5 colors). The primary y-axis (left) displays the normalized 1D sequence error (Levenshtein distance divided by N^2), contrasting the intermediate text description error (blue dotted line) against the final generated image error (red shaded line). The secondary y-axis (right) displays the 2D geometric error via normalized Spatial EMD (green solid line). Across all configurations, scaling the grid dimensions and increasing the number of categorical colors correlates with an exponential rise in spatial hallucinations and loss of geometric determinism.



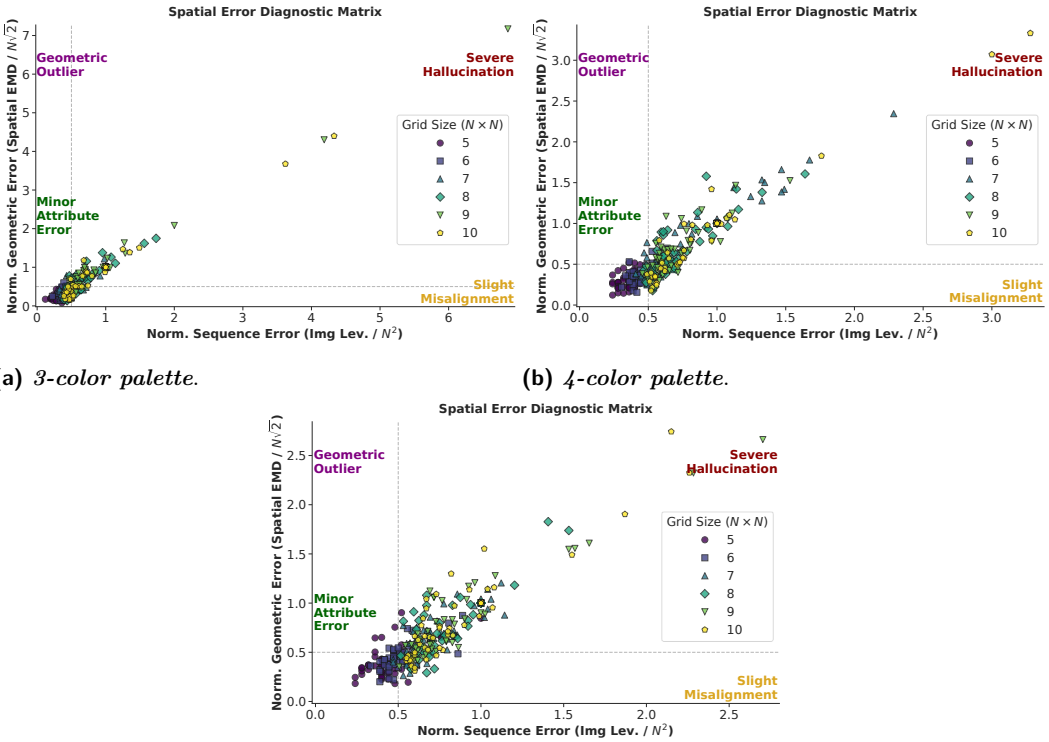
(a) *3-color palette.* **(b)** *4-color palette.*

(c) *5-color palette.*

Figure 8 Scaling behavior of average spatial reasoning errors across varying grid dimensions ($N = 5$ to $N = 10$). The left y-axis displays the normalized 1D sequence errors (Levenshtein distance divided by N^2), tracking topological failures in the text description phase (blue dashed line) and image generation phase (red dashed line). The right y-axis tracks 2D geometric error via normalized Spatial EMD (green solid line). Individual figures show different trends in these measures when performing modality transfer for cases with 3 **(a)**, 4 **(b)**, and 5 colors **(c)**.

Figure 8 shows the aggregated average measures for each grid size and color configuration to highlight the effect of increasing spatial complexity on the performance of the LMM transfer task. Overall, the average scaling trends also identify a capability drop after a certain point. For the 3-color palette (Figure 8a), this threshold occurs abruptly at $N = 7$, where the textual encodings begin to fail, cascading into significantly higher image generation errors. This phenomenon is exacerbated in the 4-color palette (Figure 8b), where the modality transfer success destabilizes earlier and more severely, resulting in a steady, proportional increase in both sequence and geometric errors as the grid size scales. At the maximum complexity of 5-color palette (Figure 8c), the models exhibit systemic failure. The initial image-to-text extraction phase (Txt Levenshtein) incurs massive errors at grid size at $N = 7$. As the foundational spatial description is intrinsically corrupted, the subsequent text-to-image generation is fed hallucinated or highly disjointed spatial coordinates. This results in maximum-penalty geometric displacements (Spatial EMD) peaking consistently across $N \geq 7$, demonstrating that the models entirely lost the capacity to deterministically map or reconstruct the original spatial context.

Figure 9 places the individual cases into four main evaluation categories introduced in Table 1, namely the Minor Attribute Error, Slight Misalignment (Topology Failure), Severe Hallucination (Catastrophic Failure), and Geometric Outlier (Shape Distortion). Each instance of an input image – generated text – and generated image (e.g., $5 \times 5 \times 3$ – img_1) is plotted as a single point using the Normalizes Sequence Error and Normalized Geometric



(a) *3-color palette.*

(b) *4-color palette.*

(c) *5-color palette.*

Figure 9 Diagnostic scatter matrices mapping 1D sequence error (Normalized Levenshtein, x-axis) against 2D geometric error (Normalized Spatial EMD, y-axis). The spatial distribution patterns directly correspond to the diagnostic categories defined in Table 1. Individual figures show diagnostic matrices for different cases with 3 **(a)**, 4 **(b)**, and 5 colors **(c)**.

Error as the axes. For the 3-color palette, the majority of points cluster in the bottom-left quadrant (Minor Attribute Error), indicating successful modality transfer at smaller grid scales. The 4-color palette (Figure 9b) induces broader dispersion, with a noticeable migration of $N \geq 8$ data points into the right hemisphere (Slight Misalignment), reflecting topological failures where sequences are corrupted despite partially intact geometries. At 5 colors (Figure 9c), the distribution is severely skewed toward the top-right quadrant (Severe Hallucination). This dense clustering highlights a catastrophic failure in modality transfer, where the model simultaneously loses both sequential topology and geometric precision. Overall, the progression visually demonstrates the LMM's failure trajectory as both grid scale and categorical color density overwhelm its spatial reasoning capabilities.

5 Discussion

While the overall results of the study are presented in Section 4, some interesting insights that were observed during the experiments are not visible from the figures and information presented there. In the following, we discuss our observations and thoughts regarding the proposed approach and the LMM modality transfer. Namely, we discuss their issues in performing image-to-text prompts and the anomalies we observed in the results of text-to-image prompts. We also provide further interpretation of the diagnostic analysis of modality transfer and lay out the limitations of our approach and case study.

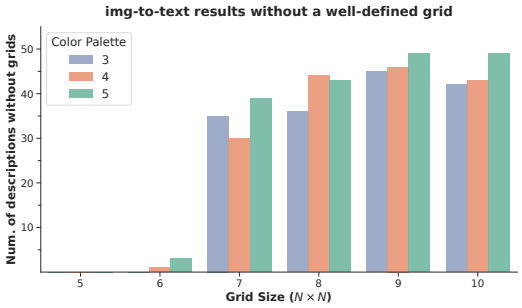


Figure 10 Bars show the number of invalid Prompt 1 (img-to-text) responses where the explicitly defined grid of colors is missing from the generated textual description. Each bar shows this value for one $N \times N \times N_{colors}$ configuration where a total of 50 prompts were executed.

5.1 Issues in Prompt 1 (Image-to-Text) outputs

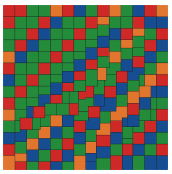
We intentionally defined the instructions to LMM in Prompt 1 loosely, not to affect the spatial comprehension of the LMM with our prompt engineering. That said, the outputs for our initial input images of 5×5 grids with 3 colors seemed quite uniform as the model would briefly describe the overall task in natural language and then provide some kind of matrix representation of the grid in the input image – often with single letters representing different colors (see the example response in Figure 4). However, when the complexity of the input grid would increase through higher cell or color count, we noticed that textual responses would sometimes not include the full grid representation at all and sometimes only include the first few rows with the note such as *"If you need a lossless, per-cell matrix for this exact image, ask for per-cell encoding and I will return a row-major 2D array"*. We noticed that this effect started occuring more often from the $N = 7$ and upwards – Figure 10 shows the number of Prompt 1 responses per configuration (out of 50) where the explicit grid is missing.

If we compare this with the evaluation of the higher dimension grid cases ($N = 8, 9, 10$) in Figure 7, we can notice that the images generated using such textual descriptions (i.e., cases where Txt Levenshtein $= 1$ and Img Levevenshtein < 1) are still able to better represent the input image. This is because the LMM is able to use the other natural language description from the Prompt 1 output to somewhat successfully recreate the input image, even if the grid configuration and colors are not excplicitly specified.

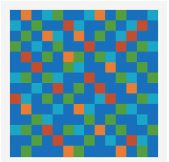
5.2 Generative Anomalies in Text-to-Image Generation

After the LMM has generated output images with Prompt 2 (text-to-image), manual processing of these images was needed to sample their grids and colors (see Section 3.3). Our goal in sampling was to rely on clearly defined criteria but also include as many grids as possible, where feasible. Thus, most grids with valid color schemes could be sampled and included in further evaluation.

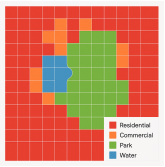
In our criteria, we assume that a valid color scheme is given when one clearly defined and closed cell (no matter what shape) is filled with a single color or uniform coloring. Therefore, grids with white or unfilled cells were also taken into account (see Figures 11f and 11g). This does not affect the evaluation, as such cases appear to be incorrect anyway. Examples where this is not the case (i.e., color scheme is invalid) are shown in Figures 11d, 11e, 11h, and 11j. Such cases were ommitted from further processing as cell colors could not be sampled and they were assigned maximum errors in the evaluation.



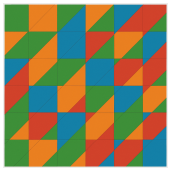
(a) target grid size: $9 \times 9 \times 4$; invalid grid with valid colors: different shapes; omitted.



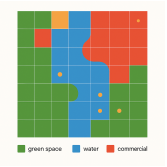
(b) target grid size: $10 \times 10 \times 4$; invalid grid with valid colors: no grid lines; included (as merged cells).



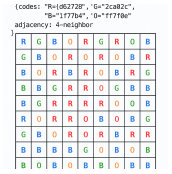
(c) target grid size: $10 \times 10 \times 4$; invalid grid with valid colors: merged and covered cells, curved grid line; omitted.



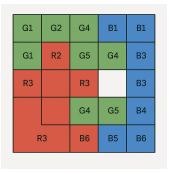
(d) target grid size: $6 \times 6 \times 4$; valid grid with invalid colors: diagonal color mix; omitted.



(e) target grid size: $7 \times 7 \times 4$; valid grid with invalid colors: river-like structure and dots; omitted.



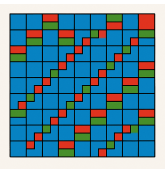
(f) target grid size: $8 \times 8 \times 4$; valid grid with invalid colors: empty/white cells; included.



(g) target grid size: $6 \times 6 \times 3$; invalid grid and colors: merged and empty/white cells; included.

(h) target grid size: $6 \times 6 \times 5$; invalid grid and colors: interrupted grid and colored symbols within cells; omitted.

(i) target grid size: $9 \times 9 \times 5$; invalid grid and colors: missing grid visualisation; omitted.



(j) target grid size: $9 \times 9 \times 5$; invalid grid and colors: different shapes and color blurring; omitted.

Figure 11 The following examples illustrate the invalid results of text-to-image generation. The target grid sizes, validity, and decision of inclusion or omittance in the process of image sampling (manual color capturing) are indicated below each image.

Borderline cases where grids are cut off (see Figure 11f) have been included, as well as grids with merged cells that follow the lines of the otherwise valid rectangular grid (see Figure 11g), as long as cells are closed and clearly defined. In contrast to grids with merged cells of other shapes like shown in Figure 11c or grids with interrupted grid lines as shown in Figure 11h. If a grid appears to have no visible lines as shown in Figure 11b it is a borderline case as well. However, grids that do not have neighbouring cells of equal color and that do not have grid lines are still sampled as valid grids. Nevertheless, the proposed solution is to treat neighbouring cells with the same color as merged cells in order to reflect this in the sampled data. Merged cells are generally sampled once, with the cells being captured in the direction from top left to bottom right at the time of their initial appearance. Grids like shown in Figure 11g are therefore treated as invalid in the evaluation and receive higher error scores. However, grids with irregularly shaped cells could not be taken into account within the sampling process, as no clear division into rows and columns is possible here. Figure 11a shows an example of an invalid grid with valid colors which had to be omitted for further evaluation.

Clearly, grids lacking meaningful semantic geometry, valid grid structures, or consistent coloring could not be encoded into discrete 2D grids. As these severe generative failures represent a complete loss of spatial determinism, they were deemed entirely invalid and had to be omitted from the quantitative analysis (see Figures 11i and 11j).

5.3 Diagnostic Analysis of Modality Transfer

By analysing the LMM performance through the lens of Spatial Error Diagnostic Matrix (Figure 9, and Table 1), we successfully categorize the nature of these breakdowns. In low-complexity scenarios (Figure 9a), the grid samples mostly aggregate in the bottom left quadrant (**Minor Attribute Eorror**, Low Lev. and Low EMD). This establishes an acceptable baseline understanding of spatial arrangement, committing several minor localized color substitutions while preserving the overarching geometric shape.

As the grid and color complexity increases, in a 4-color palette (Figure 9b), the grid samples disperse toward the first (High Lev. and Low EMD) and fourth (High Lev. and High EMD) quadrants, implying a failure of topology and showing hints of model hallucination. In this stage, while the LMMs can grasp the general colors of grid but fail to anchor them into correct, absolute coordinate geometries. The transition to the 5-color palette (Figure 9c) shows a descent into catastrophic spatial failure. The dense grid clustering extending into the top-right quadrant (**Severe Hallucination**, High Lev. and High EMD) indicates a complete collapse of cross-modal translation. In these cases, the models fail to preserve both the 1D sequential order and the 2D geometric boundaries, resulting in images where blocks are either entirely omitted, scattered randomly, or drawn using out-of-bounds categorical colors.

5.4 Limitations

While our experimental design establishes a quantifiable baseline for spatial modality transfer, several limitations need to be discussed. Currently, our empirical evaluation is limited to a narrow selection of proprietary flagship models, specifically recent iterations from OpenAI (`gpt-5` and `gpt-image-1`). While there is little reason to expect another LMM would suddenly perform this task perfectly, other models need to be tested to gain a wholistic overview of the current state-of-the-art. Additionally, the prompts in this study were executed strictly under zero-shot conditions, leaving the potential performance gains of few-shot prompting or domain-specific fine-tuning unexplored. It is thus possible that a few-shot performance would show significant improvements, but it is questionable whether it would allow for a seamless experience of working with GIS agents if every communication step between the user and the agent has to be repeated several times – at least when it involves different modalities.

In addition to being tested in a zero-short setting, our prompts have intentionally only vaguely defined the task at hand. This has possibly led to some of the observed issues such as incomplete or incorrectly structured grids. The model would likely perform better with well defined prompts where the valid output format and execution steps are explicitly defined. We believe that future research on agentic GeoAI calls for a systematic investigation of the effects of vague versus well-structured prompts on the agents' success.

Furthermore, although the procedurally generated $N \times N$ categorical grids offer a grounded and robust baseline, they represent a highly abstracted and relatively simplistic spatial reasoning task. This works well to expose the fundamental issue of current LMMs which lose spatial information through modality transfer, but is not a good representation of a full-scale GIS workflow. Because foundation models are trained on real-world data (i.e., photographs as well as textual descriptions of spatial scenes) and not grids of colored squares, they seem to struggle to generalize from context-rich real world vision-reasoning tasks [2] to simpler tasks like the one presented here. Thus, it remains to be investigated if LMMs would perform better if they were tested on more complex and well contextualized workflows embedded into a full-scale GIS project environment. If the modality transfer was not performed in isolation

on simple grid images, but on the real-world land cover maps in a software like ESRI ArcGIS, irregular shapes and additional information about them would make the task more complex, but the added context may also improve the comprehension of the LMM.

Another limitation within our current approach is the GUI-based sampling procedure in the evaluation of the generated images. In its current format, this step is not scalable to larger sets of generated images and remains vulnerable to human errors and subjective interpretation of the operator. For those reasons, it would be better to replace this step with an automated computer vision approach in the future to improve reproducibility and usability. We also only provide the joint evaluation of both modality transfer steps, while it may be informative to consider them in isolation – i.e., use ground truth input text for the text-to-image prompt. This would allow for a more detailed and objective evaluation of individual modality transfers without allowing the errors in the output of the first step to influence the second modality transfer step.

6 Conclusion

This study focuses on the fundamental issue of LMMs losing geographical information simply due to transferring it from textual to image modality or vice-versa. Anecdotally, this can often be observed when an LMM is asked to generate an image based on a textual description or to textually describe an input image (e.g., a map), and the results returned are imperfect. We argue that a (near) lossless transfer of geographical information between text and image modalities is a pre-requisite for enabling the development of autonomous GIS agents that can adequately perform GIS workflows.

To this end, we propose an LMM modality transfer task where a set of input images, here regular grids of colored squares, are to be transferred from image to text modality with Prompt 1 and then, based on the textual description outputs of Prompt 1, again be transferred from text to image modality in Prompt 2. This represents a simplified example of working with a land-cover dataset in a GIS workflow. Both outputs of Prompt 1 (generated text) and Prompt 2 (generated image) are then evaluated against the input image to quantify the amount of information lost during modality transfer.

We executed the modality transfer task on two recent OpenAI models (gpt-5 for Prompt 1 and gpt-image-1 for Prompt 2) using 900 input image with increasing grid sizes (5×5 to 10×10) and number of colors (3 to 5). Our results demonstrate that, while LMMs may process vast prior core knowledge learned from large image or language corpus, they still suffer from spatial information loss and hallucination during cross-modal translation, especially on inputs with larger grids, proving that lossless modality transfer remains a task that is yet profoundly difficult for state-of-the-art LMMs.

Future work should first apply this task to a wider array of state-of-the-art LMMs from different producers. It should also consider performing this task in a few-shot manner to avoid the issue where explicit full-grid textual representations are not generated for inputs with larger grid sizes. We also discussed the possibility that tasks using real-world geographic information instead of synthetic examples as input may, even though they seem more complex at first, yield better comprehension in LMMs as they would provide more context and are more similar to the raw training data that the LMMs are trained on.

References

1 Temitope Akinboyewa, Zhenlong Li, Huan Ning, and M Naser Lessani. GIS copilot: Towards an autonomous GIS agent for spatial analysis. *International Journal of Digital Earth*, 18(1):2497489, 2025. doi:10.1080/17538947.2025.2497489.

2 An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

3 Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report. *arXiv preprint*, 2024. doi:10.48550/arXiv.2412.04604.

4 Anthony G Cohn and Robert E Blackwell. Can large language models reason about the region connection calculus? *arXiv preprint*, 2024. doi:10.48550/arXiv.2411.19589.

5 Anthony G Cohn and Robert E Blackwell. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In Benjamin Adams, Amy L. Griffin, Simon Scheider, and Grant McKenzie, editors, *16th International Conference on Spatial Information Theory (COSIT 2024)*, volume 315 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:9, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.COSIT.2024.28.

6 Eirinaios-Odysseas Gardelakos, Vasileios Kyriakopoulos, Despina-Athanasia Pantazi, Orestis-Minas Kapopoulos, Maria Tsourma, and Manolis Koubarakis. Can Large Reasoning Models Reason about Spatial Relations? In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '25, pages 81–91, New York, NY, USA, December 2025. Association for Computing Machinery. doi:10.1145/3764912.3770841.

7 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34:625–636, April 2020. doi:10.1080/13658816.2019.1684500.

8 Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science*, 39(9):1866–1903, 2025. doi:10.1080/13658816.2025.2511227.

9 Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. URL: https://api.semanticscholar.org/CorpusID:60827152.

10 Zhenlong Li, Huan Ning, Song Gao, Krzysztof Janowicz, Wenwen Li, Samantha T. Arundel, Chaowei Yang, Budhendra Bhaduri, Shaowen Wang, A. Xing Zhu, Mark Gahegan, Shashi Shekhar, Xinyue Ye, Grant McKenzie, Guido Cervone, and Michael E. Hodgson. GIScience in the era of Artificial Intelligence: a research agenda towards Autonomous GIS. *Annals of GIS*, 31:501–536, 2025. doi:10.1080/19475683.2025.2552161.

11 M. R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research and Application*, 26:340–350, October 2001. doi:10.1002/col.1049.

12 Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint*, 2023. arXiv:2304.06798.

13 Ivan Majic, Elham Naghizade, Stephan Winter, and Martin Tomko. Discovery of topological constraints on spatial object classes using a refined topological model. *Journal of Spatial Information Science*, 18:1–30, June 2019. doi:10.5311/JOSIS.2019.18.459.

14 Ivan Majic, Elham Naghizade, Stephan Winter, and Martin Tomko. There is no way! Ternary qualitative spatial reasoning for error detection in map data. *Transactions in GIS*, 25:2048–2073, August 2021. doi:10.1111/tgis.12765.

15 Ivan Majic and Elek Pafka. AwaP-IC—An Open-Source GIS Tool for Measuring Walkable Access. *Urban Science 2019, Vol. 3, Page 48*, 3:48, April 2019. doi:10.3390/urbansci3020048.

16 Ivan Majic, Zhangyu Wang, Krzysztof Janowicz, and Mina Karimi. Spatial Task-Explicity Matters in Prompting Large Multimodal Models for Spatial Planning. *GeoAI 2024 - Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 24:99–105, November 2024. doi:10.1145/3687123.3698293.

17 Manuel Melgosa. Testing CIELAB-based Color-Difference Formulas. *Color Research & Application*, 25(1):49–55, 2000. doi:10.1002/(SICI)1520-6378(200002)25:1<49::AID-COL7>3.0.CO;2-4.

18 Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. doi:10.1023/A:1026543900054.

19 Simon Scheider and Werner Kuhn. How to talk to each other via computers: Semantic interoperability as conceptual imitation. In *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation*, pages 97–122. Springer, 2015.

20 Shaolin Xie, Shang-Ling Hsu, Qihan Zhang, Yiming Gao, Cyrus Shahabi, and Ibrahim Sabek. Evaluating Intrinsic Geospatial Topological Reasoning in LLMs. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, pages 43–48, 2025. doi:10.1145/3764915.3770722.

Mapping Cycling Experience in Valence–Arousal Space: A Geometric Approach to Appraisal Heterogeneity

Ayda Grisiute[1]
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Johanna Wörle
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Panagiotis Mavros
i3 (CNRS), Department of Social and Economic Sciences, Télécom Paris, Institut Polytechnique de Paris, France

Martin Raubal
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Abstract

In this paper, we argue that heterogeneity in experiential appraisals can function as an informative signal for subjective cycling experience. We introduce a geometric framework to characterize the "shape" of experiential appraisal within a two-dimensional affect space. Using data from a crowd-sourced survey, in which participants ($N = 239$) evaluated 30 first-person cycling video scenarios, we describe appraisal heterogeneity not only by its overall extent (dispersion), but also by its directional structure (anisotropy), bimodality (polarization), and the diversity of affective categories and environmental cues (entropy). Our findings reveal a structural asymmetry. Negative environments result in a rigid, more linear stress response, whereas positive environments allow for diffuse, multi-polar appraisals. We further find that observer experience modulates the magnitude of dispersion, whereas the direction of appraisal remains relatively stable across demographic subgroups. We propose that this approach can quantitatively distinguish between those environments that are consistently perceived as non-bikeable and those that evoke conflicting interpretations. Our methodology complements user-centric models, such as cycling archetypes, by describing the underlying structure of divergent experiences.

2012 ACM Subject Classification Applied computing → Transportation; Applied computing → Psychology

Keywords and phrases bike infrastructure, affect grid, inter-rater variability, online survey

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.8

Supplementary Material *Software*: `https://github.com/mie-lab/cycling_experience`
Audiovisual: `https://www.youtube.com/@cycling-survey/playlists`

Declaration on GenAI/LLM use Generative AI tools were used for language editing and formatting. All scientific content, analyses, and conclusions were developed and validated by the authors.

1 Introduction

Cycling experience is widely recognized as a multidimensional construct, shaped not only by physical infrastructure but also by perceptual, affective, and cognitive processes [29]. As an embodied mode of travel, cycling requires the continuous processing of sensory, emotional, and interactional cues [27]. Experiential appraisals also have practical importance: perceived

[1] Corresponding author

© Ayda Grisiute, Johanna Wörle, Panagiotis Mavros, and Martin Raubal;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 8; pp. 8:1–8:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

comfort and safety may directly influence cycling uptake, with downstream impacts on public health, urban livability, and transport-related emissions [11]. In turn, rather than relying on objective infrastructural descriptors alone, researchers are increasingly measuring and modeling cycling experience through affective responses to environmental stimuli [8, 17].

Many studies report large appraisal variation which they commonly attribute to observer characteristics (demographics, archetypes) [7, 3]. This focus overlooks the fact that environments with identical mean ratings may elicit distinct response patterns. For instance, when evaluating the same street, cyclists' emotional responses might vary, appearing as a scattered cloud of data points (i.e., diffused). Alternatively, riders might split into distinct opinion clusters, such as half finding the street exciting and the other half finding it highly stressful (i.e., polarized). We refer to these structural characteristics of how appraisals are distributed across riders as *structural response properties*. However, little research has focused on analyzing these properties.

Structural differences could reveal whether an environment sends clear or conflicting signals to its users, serving as a diagnostic signal of *shared environmental legibility*. Urban legibility traditionally refers to the ease with which a city's parts can be recognized and organized into a coherent mental pattern [19]. Here, we extend this notion beyond navigational clarity to how reliably environmental cues convey a shared affective meaning. To assess this empirically, we need a representation in which structural response properties become visible. When subjective experience is represented two-dimensionally, distinct geometric signatures can emerge. Responses may be broadly scattered (dispersion), elongated along a dominant affective dimension (anisotropy), clustered into distinct camps (polarization), or spread across qualitatively distinct affective states (entropy). These properties offer a novel way to audit cycling environments, as scenarios with similar mean ratings may require different interventions depending on whether appraisals are consistent or contested.

To investigate these ideas, we re-analyzed a crowd-sourced online survey in which participants evaluated short, first-person videos of cycling scenarios using the Affect Grid, a two-dimensional measure of affective valence (pleasantness) and arousal (activation) [25]. By combining controlled video stimuli with a repeated-measures design, we obtained dense two-dimensional response distributions for each cycling scenario. Both the location and the structure of affective responses in valence–arousal space were examined to assess the extent to which this variation is driven by environmental properties as opposed to observer covariates.

We hypothesize that subjective cycling experience is not a product of cycling infrastructure (used here as an umbrella term for dedicated facilities such as separated bike lanes, advisory lanes, and shared-use paths) or observer traits alone, but that specific environments systematically elicit structured appraisal heterogeneity signatures. The following three research questions will be addressed:

- **RQ1.** How do cycling scenarios differ in valence–arousal appraisal space?
- **RQ2.** What distinct response geometry signatures characterize cycling scenarios, and how do they relate to environmental cues?
- **RQ3.** To what extent is the structure of appraisal heterogeneity moderated by observer characteristics versus scenario properties?

To evaluate this approach, we introduce a set of geometric metrics that characterize the structural response patterns of subjective appraisals (Sections 2 and 3). We demonstrate that high- and low-stress environments generate distinctly different appraisal signatures, which our geometric framework captures reliably across diverse observer backgrounds (Sections 4 and 5).

2 Background

This section introduces existing work on subjective cycling experience and discusses how it is typically measured. We first discuss cycling experience as an emerging multidimensional (perceptual, cognitive, and sensory) construct, and then evaluate current modeling and measurement approaches and sources of inter-subjective variability.

2.1 Cycling Experience as a Multidimensional Construct

Cycling research has traditionally evaluated cycling environments using objective indicators such as facility type, traffic volume, or network connectivity [8]. However, these descriptors only partially predict subjective experience [4, 5]. This leaves a "perception gap" where similar cycling infrastructures may evoke highly variable appraisals or, conversely, distinct environments result in comparable appraisals. For instance, low-traffic streets without dedicated cycling facilities are often rated similarly to high-traffic streets equipped with such facilities [1, 17]. Subjective cycling experience is therefore increasingly conceptualized as a perceptual–cognitive appraisal. In this paper, we adopt the definition of an episodic affective and evaluative response to environmental stimuli encountered during a trip from Lim et al. [18]. This process is inherently embodied: unlike car drivers or transit users, cyclists are continuously exposed to their surroundings, requiring the real-time processing of visual, auditory, and interactional signals [6, 27].

Most empirical studies attribute variability in environmental appraisals primarily to observer-related factors, such as demographic characteristics or cycling archetypes, which reflect variation in confidence, risk tolerance, habits, and cycling purpose [7, 3]. While valuable for identifying which individuals experience the environment differently, these approaches often treat observer heterogeneity as the primary cause. They may overlook the possibility that certain environments inherently afford multiple interpretations [22].

2.2 Tools and Modalities for Cycling Experience Measurement

Research on subjective cycling experience employs a range of measurement modalities, from in-situ ratings and physiological sensors to using virtual environments or videos as stimuli [16, 8]. These differ in the richness and timing of information provided, potentially capturing distinct facets of the experience, where evaluations emerge through different processes (recall, immediate reflection, sensory responses) [13].

Stimulus. When cycling, relevant cues unfold dynamically: perceived safety and comfort depend on motion, closing speeds, and overtaking dynamics [17]. While static images require viewers to infer dynamic processes, video and virtual environments (VR) preserve visual continuity and interaction cues essential for spatial affordance processing [9, 24]. Although they lack the vestibular and haptic feedback of physical cycling, research on the behavioral validity of traffic and cycling simulators suggests that sensory realism beyond a task-relevant threshold may yield diminishing returns for specific judgment tasks [26, 21].

Sampling strategy. Examining observer-driven variance vs. stimulus-driven ambiguity typically requires large, diverse samples. Although naturalistic studies offer high realism, they often rely on small, homogeneous samples. Online crowd-sourced surveys trade experimental control for access to heterogeneous participant pools [12]. In the context of cycling experience, several studies emphasize that large and diverse samples are essential to better differentiate stimulus-driven variation from observer-related differences [14, 1].

Measurement scales. Objective measures to assess cycling experience typically use psychophysiological sensors (e.g., heart rate, skin conductance) to capture variations in physiological arousal, commonly an indicator of vigilance and stress [16, 28]. Subjective measures, on the other hand, rely on self-report methods such as interviews or rating scales to measure overall appraisal (e.g., comfort, safety, pleasantness), thus primarily reflecting affective valence [28]. However, both dimensions are essential for a holistic cycling evaluation. Affect research suggests that the valence–arousal relationship depends on stimulus clarity. In clear situations, pleasantness and stress covary in a direct, predictable manner. In contrast, when environments present ambiguous affective cues, pleasantness and stress do not predictably covary, resulting in scattered responses that vary widely across individuals [20, 15]. Yet, cycling research has rarely explored whether mixed reactions to a single environment reflect the cyclists' characteristics, or if certain spaces naturally elicit polarized responses.

Two instruments are most commonly used in the joint measurement of valence and arousal: the *Self-Assessment Manikin* (SAM) [2] and Russell's *Affect Grid* [25]. SAM uses three independent pictorial scales for valence, arousal, and dominance, sampled sequentially. The Affect Grid captures valence and arousal jointly as a single choice in a 2D coordinate space, omitting dominance. In this study, we adopt the Affect Grid for two reasons. First, its single-choice, two-dimensional mapping minimizes cognitive fatigue during the rapid evaluation of 30 scenarios, whereas SAM requires three sequential ratings per stimulus. Second, the Affect Grid is inherently spatial, forming a 2D coordinate system that directly supports the geometric metrics central to our analysis.

3 Methodology

We developed this geometric approach to subjective experience by examining response distributions in valence–arousal space. Our methodology combined (i) first-person video stimuli of diverse cycling scenarios and (ii) a repeated-measures online survey that yielded dense affective response distributions across participants. We quantified not only mean affect but also the magnitude and structure of appraisal. Finally, we tested whether observer covariates systematically moderated responses to high-heterogeneity cycling scenarios. Figure 1 provides an overview of the study methodology.

3.1 Study Design

Participants, Recruitment, and Ethics. Participants ($N = 313$) were recruited via institutional mailing lists, social media, physical leaflets, and newsletters using snowball sampling. Inclusion criteria required age ≥ 18, normal or corrected-to-normal vision, and sufficient English proficiency. Participation was voluntary and could be discontinued at any time. Data were stored, processed, and reported in accordance with ETH Zurich data protection practices. Ethical approval was granted by the institutional ethics committee (Application 24-ETHICS-386).

Cycling Scenario Generation and Selection. First-person perspective cycling footage was recorded in Zurich, Switzerland between September and October 2024. To minimize confounding effects from extreme congestion or weather, recording occurred during fair weather and non-peak hours. Footage was captured using a handlebar-mounted action camera (DJI Action 4) on an e-bike, maintaining a mean speed of 16.5 km/h, with small variations due to local topography. Raw footage was processed to remove intersections, stops, and maneuvering, keeping linear segments which were then cut into 30-second clips. From this

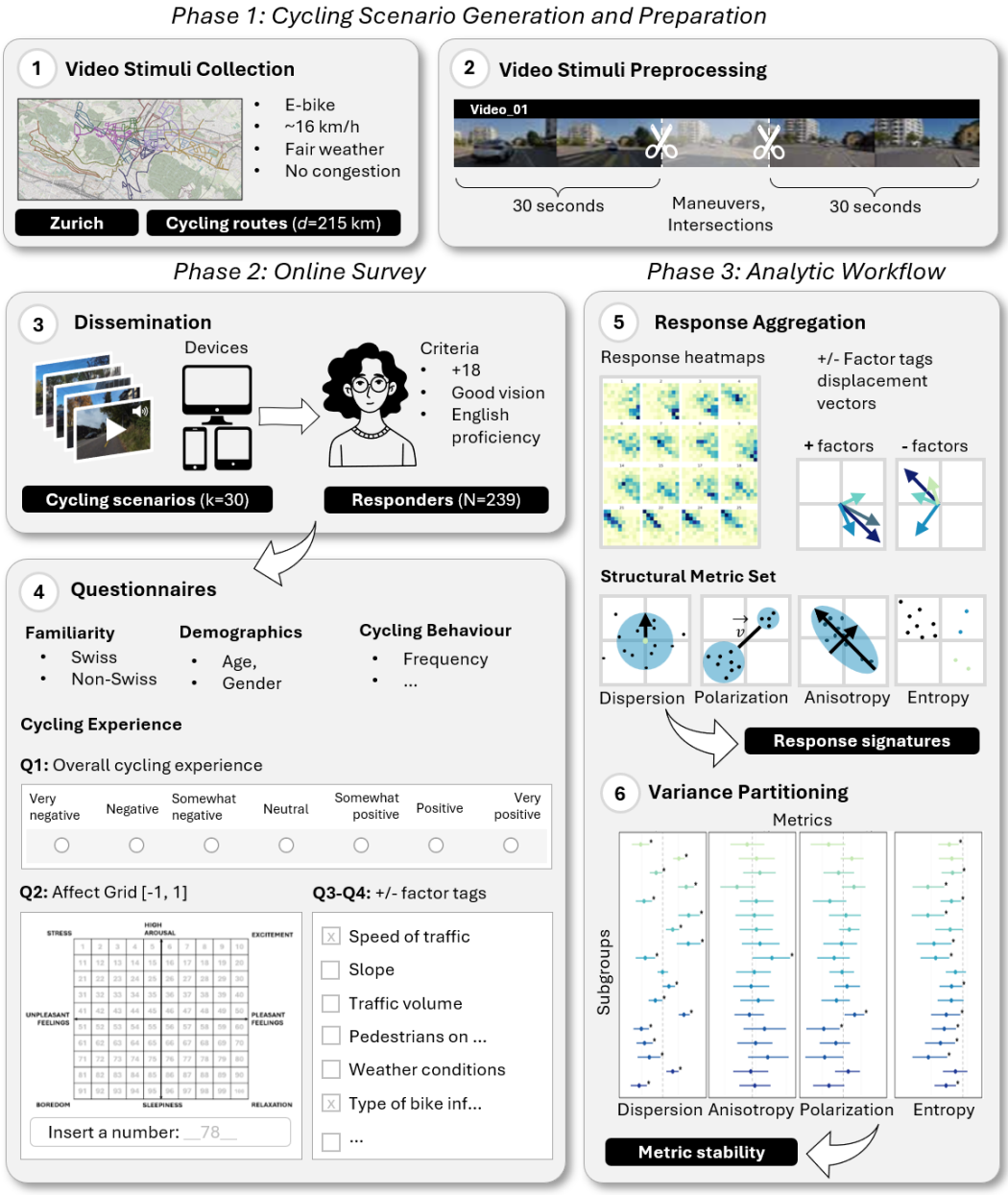


Figure 1 Procedural workflow of the study methodology.

pool, 30 scenarios (see Supplementary Material) were selected to sample diverse co-occurring environmental characteristics drawn from the cycling literature [1, 10]. Specifically, these factors include cycling infrastructure type, traffic speed and volume, car overtakes, slope, surface quality, pedestrian and cyclist presence, surrounding greenery and buildings, and weather conditions (see Figure 2). They correspond with the positive and negative factors used in our questionnaire (Section 3). Additional factors (e.g., public transit lanes, points of interest, side parking) were excluded to keep the questionnaire workload manageable and avoid respondent fatigue. Privacy was ensured by anonymizing faces and license plates using the EgoBlur model followed by manual verification [23].

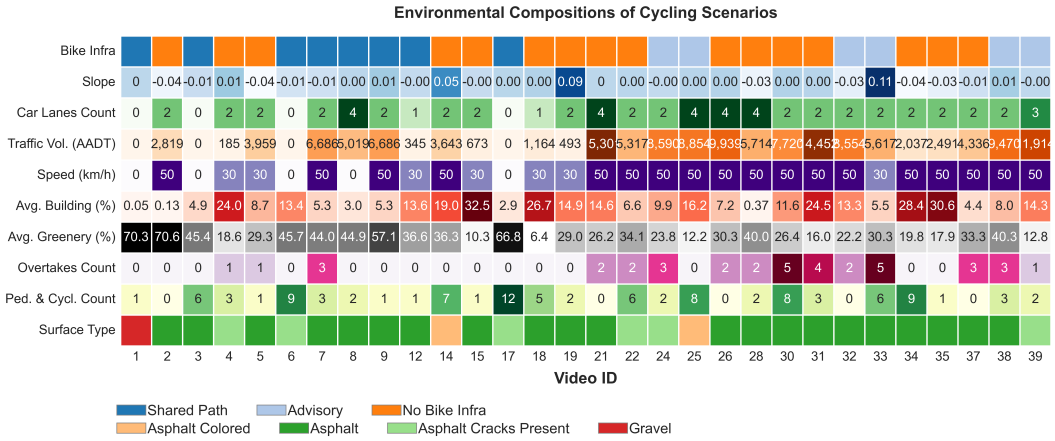


Figure 2 The co-occurrence and variance of specific infrastructural and environmental attributes across the selected video scenarios.

Procedure. The survey was implemented in Microsoft Forms and administered in English. After providing informed consent, participants completed a background questionnaire capturing demographics and cycling characteristics (purpose, environment, frequency, confidence). Participants then viewed the 30 video clips (with available audio) in randomized order. Following each scenario, they provided three evaluations: (i) an overall cycling experience rating on a 7-point rating scale to serve as a unidimensional benchmark; (ii) an affective appraisal using the Affect Grid [25]; and (iii) a selection of perceived positive and negative environmental factors introduced above. We used country of residence (Switzerland versus other) as a high-level proxy for environmental familiarity to test for moderators in affective responses. The Affect Grid required participants to choose a single cell from a valence–arousal coordinate system organized into four quadrants, which was then mapped onto a continuous $[-1, 1]$ scale for geometric analysis (see Figure 1). The cycling scenario viewing time was 15 minutes with an expected average completion time of 30 minutes.

3.2 Geometric Characterization of Appraisal

We treated the set of valence–arousal responses for each cycling scenario as a distinct geometric object in two-dimensional affect space. This step was motivated by the fact that it is often overlooked in the literature and by the observation from an initial visual screening that scenarios with similar mean ratings can arise from qualitatively different response configurations: some may cluster tightly, while others spread diffusely, align along a dominant affective axis, or split into competing clusters. To reveal these patterns, we quantified appraisal structure using five metrics (see Table 1). These metrics separated the magnitude of heterogeneity (Dispersion) from its structural properties: its directional orientation (Anisotropy Index) and bimodality (Polarization Index). We also utilized Quadrant Entropy and Cue Entropy to assess whether affective ambiguity aligned with a diversity of perceived environmental cues. Our approach is based on the premise that the valence–arousal relationship is conditional on stimulus clarity [20].

Table 1 Taxonomy of metrics for quantifying appraisal heterogeneity in the $[-1, 1]$ valence (V) and arousal (A) space. Indices j and i denote cycling scenario and participant; N_j is the sample size for scenario j.

Metric	Sym.	Definition, Formula & Diagnostic Interpretation
Location & Magnitude		
Affective Centroid	$\mathbf{C}_j$	*Definitions:* The arithmetic mean (center of mass) of the affective responses. $\mathbf{C}_j = (\bar{V}_j, \bar{A}_j)$ *Interpretation:* The aggregate affective "consensus" for scenario j.
Dispersion	D_j	*Definitions:* The average Euclidean distance of individual responses from the scenario's centroid. $D_j = \frac{1}{N_j} \sum_i \|\mathbf{r}_{i,j} - \mathbf{C}_j\|_2$ *Interpretation:* The magnitude of appraisal heterogeneity, with higher values (larger radius) indicating lower consensus. In the $[-1, 1]$ Affect Grid coordinate system, the maximum possible dispersion is $D \approx 2.83$ (extreme cases in opposite corners); a range of 0.59–0.77 represents moderate spread relative to the scale's bounds.
Geometric Structure		
Anisotropy Index	AI_j	*Definitions:* A measure of the directional bias of the response distribution, derived from the eigenvalues ($\lambda_{1,2}$) of the 2×2 covariance matrix. $AI_j = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$ *Interpretation:* Response "shape." $0 \rightarrow$ isotropic (diffuse); $1 \rightarrow$ linear (structured coupling).
Polarization Index	PI_j	*Definitions:* The Euclidean distance between the mean vectors ($\boldsymbol{\mu}$) of a 2-component Gaussian Mixture Model (GMM). Note: $PI_j > 0$ only if $\Delta\text{BIC} \geq 6$ (conservative threshold) and cluster separation ≥ 0.5. $PI_j = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2$ *Interpretation:* The structural splitting. $0 \rightarrow$ unimodal consensus; $0.8 - 1.0 \rightarrow$ moderate polarization (likely adjacent quadrants); $1 \rightarrow$ presence of distinct clusters (likely opposing quadrants).
Categorical Ambiguity		
Quadrant Entropy	H_j^Q	*Definitions:* The Shannon entropy of the response distribution across the four affective quadrants (e.g., Stress, Excitement), where p_k is the relative count in quadrant k. $H_j^Q = -\sum_{k=1}^{4} p_k \ln p_k$ *Interpretation:* Categorical conflict. High values indicate that responses span multiple affective states rather than a single quadrant.
Cue Entropy	H_j^{cue}	*Definitions:* The Shannon entropy of tagged environmental factors, where p_m is the proportion of factors falling into category m (positive or negative). $H_j^{\text{cue}} = -\sum p_m \ln p_m$ *Interpretation:* The diversity of environmental attribution, which measures the consistency of the environmental signal indicated by respondents.

3.3 Analysis Workflow

We operationalized affective response distributions as geometric objects to quantify: (i) the affective position of cycling scenarios; (ii) the magnitude and structural organization of response heterogeneity; and (iii) the stability of these patterns across observer covariates. Prior to analysis, responses were screened for incomplete submissions and completion times that were too short (< 15 minutes). Validated responses were projected into a continuous 2D coordinate system $[-1, 1]$ for analysis. Several demographic variables were aggregated into broader categories to ensure sufficient subgroup sizes for stable estimation (e.g., cycling frequency into Frequent vs. Infrequent). To establish a baseline of within-subgroup agreement in scenario-specific appraisals, we computed the intraclass correlation coefficient ICC(2,1), separately for valence and arousal across the 30 scenarios. This two-way random-effects coefficient estimates the reliability of single ratings, quantifying the expected agreement between two randomly selected participants from the same subgroup, and penalizes both random disagreement and systematic differences in scale use (rater bias). The workflow proceeded in three stages, following the order of the study's research questions.

Stage 1: Mapping the affective response landscape

We first established a global baseline by aggregating all responses to characterize the overall utilization of the valence–arousal space in the context of cycling scenarios. We then characterized each scenario by its affective centroid $\mathbf{C}_j = (\bar{V}_j, \bar{A}_j)$, providing a descriptive overview of how environments are distributed across the affect space. For each scenario, we applied Principal Component Analysis (PCA) to the scatter of individual ratings to identify the dominant axis along which individual responses for that scenario vary.

Stage 2: Diagnosing geometric appraisal signatures

We quantified appraisal heterogeneity at the scenario level using geometric metrics capturing Dispersion (D_j), Anisotropy (AI_j), and Polarization (PI_j). For each scenario, PI_j was computed by fitting one- and two-component Gaussian Mixture Models (GMMs) to the valence–arousal responses and comparing them via BIC. If a two-component model was preferred, PI_j was set to the Euclidean distance between the means of the two clusters; otherwise, it was set to zero. GMMs used full covariance matrices and 10 random initializations for stability. To distinguish between diffuse stochastic variance and structured signals, we examined the relationships among these metrics and performed split-sample reliability tests across random and subgroup partitions. To examine the environmental context of this heterogeneity, we analyzed the association between Quadrant Entropy (H_j^Q) and Cue Entropy (H_j^{cue}), testing whether conflicting affective states co-occur with the diverse perception of environmental cues. To further characterize the role of these cues, we calculated conditional displacement vectors, where each vector represents the affective shift associated with a specific tagged environmental factor. For each scenario, we computed the mean valence–arousal response of the subset of participants who tagged that factor (as positive or negative) and subtracted the scenario centroid, then aggregated these centered shifts across scenarios to yield one characteristic vector per factor.

Stage 3: Testing subgroup robustness of appraisal structure

We evaluated whether the observed appraisal structure reflected stable properties of the cycling scenarios or varied systematically across observer subgroups. To distinguish scenario-driven signals from observer-driven variance, we recomputed the scenario-level metrics

Table 2 Mean ratings for valence ($\bar{V}$) and arousal ($\bar{A}$), calculated from raw, unaveraged individual responses, alongside inter-rater agreement within demographic and cycling-related subgroups across the 30 scenarios. Higher values indicate stronger collective within-subgroup agreement on scenario-specific ratings.

Category	Count (n)	%	$\bar{V}$	$\bar{A}$	$\bar{V}$ ICC(2,1)	$\bar{A}$ ICC(2,1)
Gender						
Female	122	51.0	0.00	0.12	0.281	0.177
Male	117	49.0	0.04	0.14	0.321	0.203
Age						
18–25 years	92	38.5	0.07	0.05***	0.209	0.159
26–35 years	82	34.3	-0.01	0.17	0.390	0.262
36–45 years	39	16.3	0.01	0.16	0.340	0.201
46+ years	26	10.9	-0.03	0.25	0.342	0.114
Cycling Frequency						
Infrequent	126	52.7	0.03	0.11	0.231	0.148
Frequent	113	47.3	-0.03	0.16	0.398	0.246
Cycling Purpose						
Commuting	114	47.7	0.03	0.14	0.390	0.246
Recreational	98	41.0	0.02	0.12	0.234	0.135
I do not cycle	27	11.3	-0.05	0.14	0.228	0.186
Cycling Confidence						
Confident	196	82.0	0.03	0.13	0.326	0.209
Not confident	43	18.0	-0.02	0.14	0.200	0.120
Cycling Environment						
Urban area	162	67.8	0.06*	0.14	0.304	0.190
Rural area	49	20.5	-0.05	0.10	0.314	0.208
I do not cycle	25	10.5	-0.05	0.14	0.250	0.171
Both	3	1.3	-0.12	0.22	–	–
Familiarity						
Swiss	87	36.4	0.02	0.13	0.422	0.270
Non-Swiss	152	63.6	0.02	0.13	0.246	0.152

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (post-hoc tests).

(D_j, AI_j, PI_j, H_j^Q, and H_j^{cue}) separately for each participant subgroup (cycling frequency, confidence, purpose, gender, age, and environmental familiarity). Uncertainty was estimated using $n = 1000$ bootstrap resamples over the scenarios to calculate 95% confidence intervals. If subgroup estimates tracked the overall (all-participant) values with substantial overlap in confidence intervals, it would indicate that the appraisal "shape" is an intrinsic, scenario-specific property. On the contrary, systematic shifts and little overlap in confidence intervals would suggest that appraisal geometry is primarily moderated by observer background.

4 Results

This section presents the findings from our online survey, structured around the three research questions (see Section 1). We first describe sample characteristics, then analyze scenario-specific appraisal signatures based on the defined set of metrics. Finally, we partition the observed variance to determine whether the geometric structure of these appraisals is driven by the cycling scenario properties or moderated by observer characteristics.

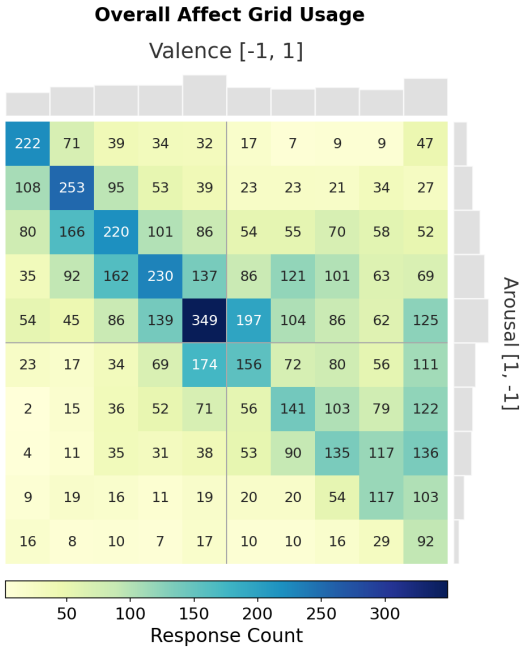


Figure 3 Overall usage of the Affect Grid across all cycling scenarios. Cell intensity indicates how frequently each valence–arousal cell was selected.

4.1 Sample, Measurement Scale Coverage, and Baseline Agreement

After quality checks, $N = 239$ valid responses were retained. Overall, mean valence was clustered around neutrality (≈ 0), while mean arousal was consistently positive, reflecting the vigilant nature of cycling (see Table 2). Arousal baselines vary by age, with the youngest cohort (18–25) reporting significantly lower arousal than older groups, which might reflect

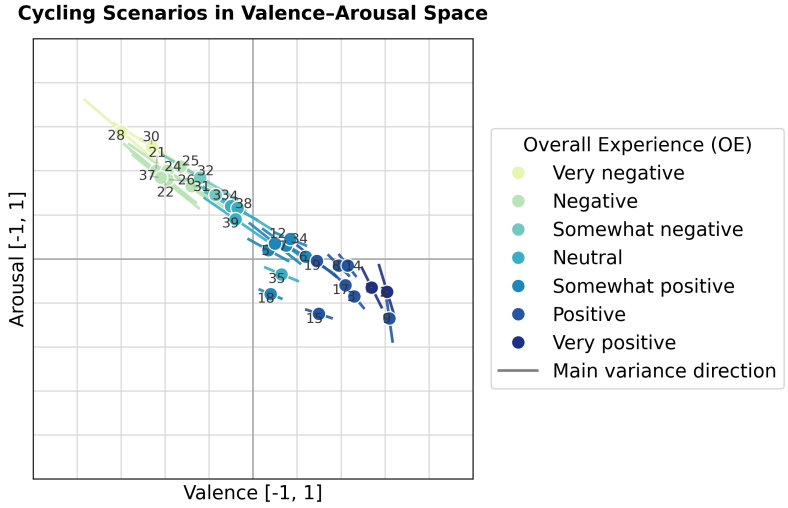


Figure 4 Cycling scenarios in valence–arousal space. Points indicate cycling scenario centroids (means), colored by modal Overall Experience (OE) rating. Line segments represent the main axes along which response varies (first principal component of the valence-arousal covariance matrix).

Table 3 Scenario-level geometric signatures. $(\bar{V}, \bar{A})$ and SD represent the affective centroid and 1-dimensional spread. ρ_{VA} denotes the valence–arousal correlation; D is the 2-dimensional mean dispersion; AI is anisotropy (directional structure); PI is polarization (clustering); and H^Q is quadrant entropy (categorical ambiguity). OE is the modal Overall Experience rating.

Scenario	$\bar{V} \pm SD$	$\bar{A} \pm SD$	OE (Mode)	ρ_{VA}	D	AI	PI	H^Q
1	0.54 ± 0.55	-0.13 ± 0.64	Very positive	-0.20	0.77	0.27	1.11	1.12
2	0.61 ± 0.46	-0.15 ± 0.63	Very positive	-0.27	0.71	0.37	1.17	1.05
3	0.46 ± 0.50	-0.17 ± 0.52	Positive	-0.29	0.66	0.21	0.94	1.13
4	-0.10 ± 0.55	0.24 ± 0.45	Neutral	-0.35	0.63	0.39	0.92	1.22
5	0.07 ± 0.57	0.04 ± 0.49	Somewhat positive	-0.27	0.69	0.30	1.03	1.35
6	0.24 ± 0.61	0.01 ± 0.51	Somewhat positive	-0.31	0.74	0.39	1.22	1.25
7	0.15 ± 0.54	0.07 ± 0.48	Somewhat positive	-0.35	0.67	0.37	0.92	1.33
8	0.39 ± 0.52	-0.03 ± 0.52	Positive	-0.18	0.68	0.18	0.95	1.24
9	0.62 ± 0.41	-0.27 ± 0.56	Positive	-0.15	0.62	0.32	0.85	0.95
12	0.10 ± 0.59	0.07 ± 0.54	Somewhat positive	-0.39	0.74	0.41	1.10	1.32
14	0.43 ± 0.48	-0.03 ± 0.50	Positive	-0.21	0.63	0.16	0.88	1.18
15	0.30 ± 0.57	-0.25 ± 0.50	Positive	-0.12	0.68	0.18	1.04	1.22
17	0.42 ± 0.54	-0.12 ± 0.53	Positive	-0.25	0.70	0.22	0.93	1.23
18	0.08 ± 0.56	-0.16 ± 0.50	Somewhat positive	-0.09	0.69	0.16	0.89	1.35
19	0.29 ± 0.55	-0.01 ± 0.52	Positive	-0.31	0.70	0.30	0.97	1.29
21	-0.44 ± 0.52	0.40 ± 0.47	Negative	-0.62	0.59	0.55	1.10	0.84
22	-0.40 ± 0.54	0.36 ± 0.49	Negative	-0.58	0.60	0.56	0.86	0.87
24	-0.39 ± 0.57	0.40 ± 0.46	Negative	-0.54	0.60	0.60	1.12	0.84
25	-0.33 ± 0.60	0.42 ± 0.48	Negative	-0.62	0.66	0.64	1.06	0.95
26	-0.39 ± 0.56	0.36 ± 0.50	Negative	-0.47	0.62	0.48	1.17	0.87
28	-0.60 ± 0.55	0.58 ± 0.50	Very negative	-0.63	0.61	0.60	0.94	0.73
30	-0.46 ± 0.58	0.51 ± 0.45	Very negative	-0.60	0.61	0.57	1.08	0.77
31	-0.28 ± 0.57	0.33 ± 0.45	Negative	-0.56	0.62	0.53	0.94	1.00
32	-0.24 ± 0.60	0.37 ± 0.46	Somewhat negative	-0.51	0.65	0.56	1.11	1.00
33	-0.17 ± 0.56	0.29 ± 0.43	Somewhat negative	-0.42	0.61	0.45	0.94	1.09
34	0.17 ± 0.55	0.09 ± 0.46	Somewhat positive	-0.18	0.64	0.23	0.00	1.32
35	0.13 ± 0.60	-0.07 ± 0.51	Neutral	-0.19	0.72	0.24	0.00	1.36
37	-0.42 ± 0.52	0.37 ± 0.47	Negative	-0.50	0.59	0.46	0.98	0.92
38	-0.07 ± 0.57	0.23 ± 0.45	Neutral	-0.48	0.65	0.51	0.95	1.21
39	-0.08 ± 0.58	0.18 ± 0.49	Neutral	-0.45	0.68	0.47	0.92	1.25

higher risk tolerance. Urban cyclists report significantly higher valence than rural cyclists. We see few differences in mean affect across gender, frequency, or confidence. However, utility-oriented traits (commuting as primary purpose, high confidence and frequency) and environmental familiarity (Swiss vs. non-Swiss) are associated with higher collective agreement ($ICC(2,1) \approx 0.33 - 0.42$), though differences are not significant. This suggests that while experience and familiarity do not alter the nature of the affective response (i.e., the means remain stable), they may drive more aligned expectations for cycling infrastructure. Figure 3 shows that overall grid usage is anisotropic, with responses clustering along a diagonal from calm–pleasant to tense–unpleasant. The two affective dimensions exhibit a moderate negative correlation ($r = -0.49$). The low-valence/low-arousal quadrant is rarely selected, likely due to the cycling task being inherently active and vigilant. An additional factor may be the stimulus design: the 30-second video clips did not include static congestion and extended delays, which would be more likely to elicit states of boredom or fatigue.

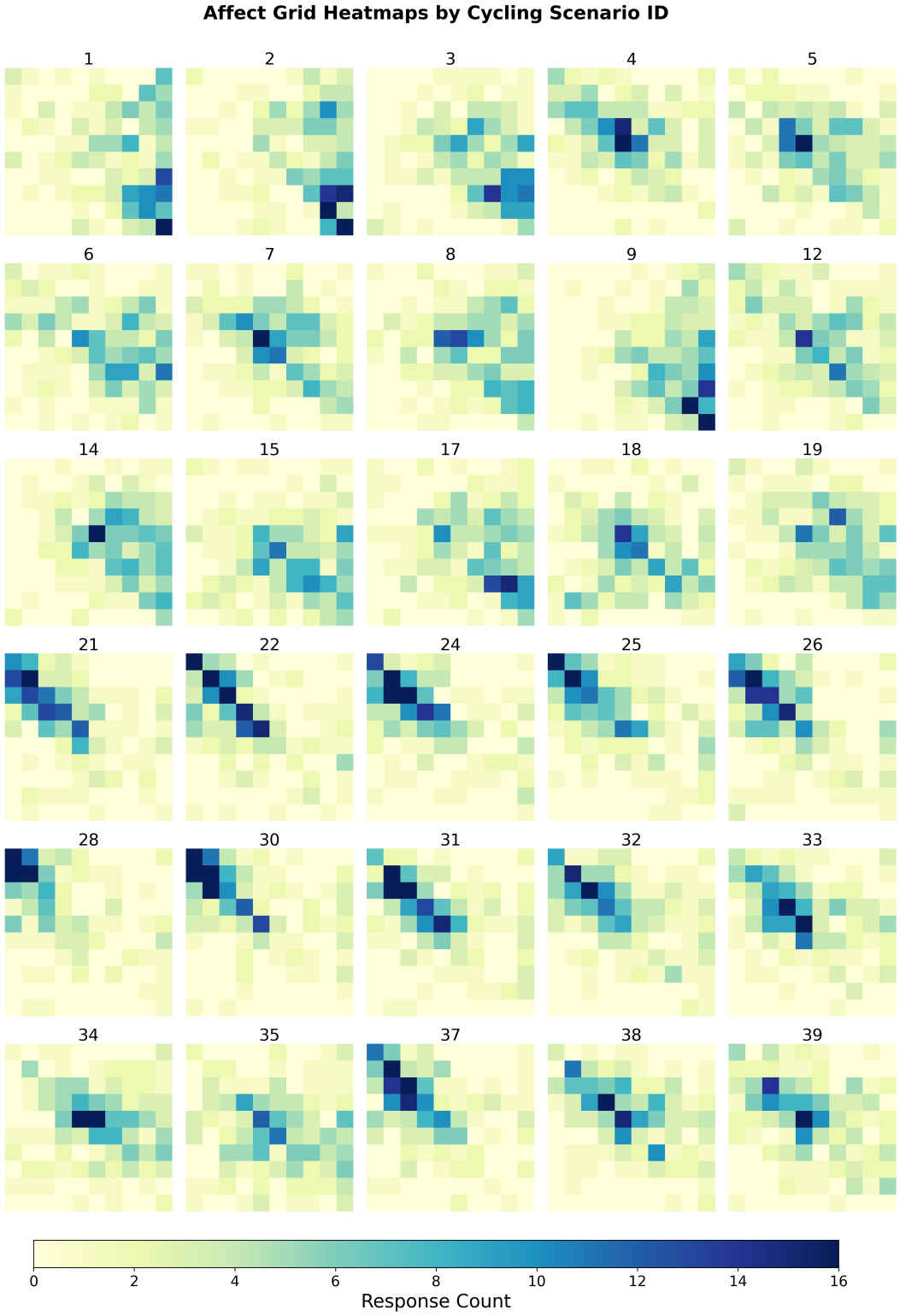


Figure 5 Response distribution across 30 cycling scenarios. Videos for all street scenarios are available online (see Supplementary Material). Darker cells indicate higher concentrations of responses for that specific scenario.

4.2 Stage 1: Mapping the Affective Cycling Experience (RQ1)

Scenario centroids and their dominant axes of variation reveal distinct structural response properties (see Figure 4). While responses generally align along the valence–arousal diagonal, the geometry of consensus differs across cycling scenarios. Negatively rated scenarios (e.g., 22, 28, 30) manifest as tight, linear streaks anchored in the high-arousal quadrant. In contrast, positively rated (e.g., 1, 2, 9) and neutrally rated scenarios (e.g., 12, 18, 35) appear as diffuse clouds, suggesting that while stress elicits uniform appraisal, positive experiences elicit more heterogeneous interpretations (see Figure 5). In stressful environments, a clear signal tightly couples valence and arousal along a shared axis, whereas in positive environments, these dimensions vary more independently. This complements the broader findings of Mattek et al. [20], who showed that the correlation between valence and arousal weakens as stimulus ambiguity increases.

4.3 Stage 2: Appraisal Signatures in Affect Space (RQ2)

Geometric Structure of Affective Responses. Table 3 quantifies the internal structure of appraisal across scenarios. While total dispersion (D) remains relatively stable (0.59–0.77), other metrics vary significantly. Negatively rated scenarios (e.g., 25, 28) show high anisotropy ($AI \geq 0.60$) and strong negative valence–arousal correlations ($\rho_{VA} \approx -0.60$), indicating that responses align tightly along a single affective axis. Positive and neutral scenarios show lower anisotropy, with responses spread more evenly across both affect axes. The polarization index (PI) further isolates specific cases, such as Scenario 6 ($PI = 1.22$), where moderate dispersion ($D = 0.74$) reflects not a shared moderate experience but a split into distinct clusters (e.g., stress vs. excitement) that the centroid alone would mask.

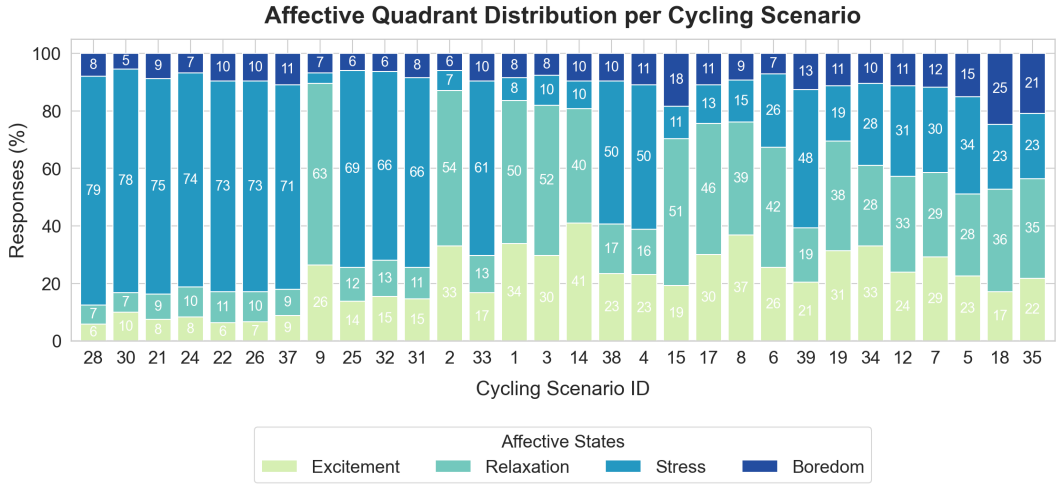


Figure 6 Response distributions across four affective quadrants for each cycling scenario. Scenarios are arranged from left to right in ascending order of Quadrant Entropy (H^Q), with the lowest-entropy (most consensus) scenarios on the left and the highest-entropy (most ambiguous) scenarios on the right.

Environmental Drivers and Directional Shifts. As shown in Figure 6, neutral scenarios exhibit high Quadrant Entropy (H^Q) and span multiple affective quadrants, whereas negative scenarios are dominated by the "Stress" quadrant. H^Q correlates moderately with H^{cue}

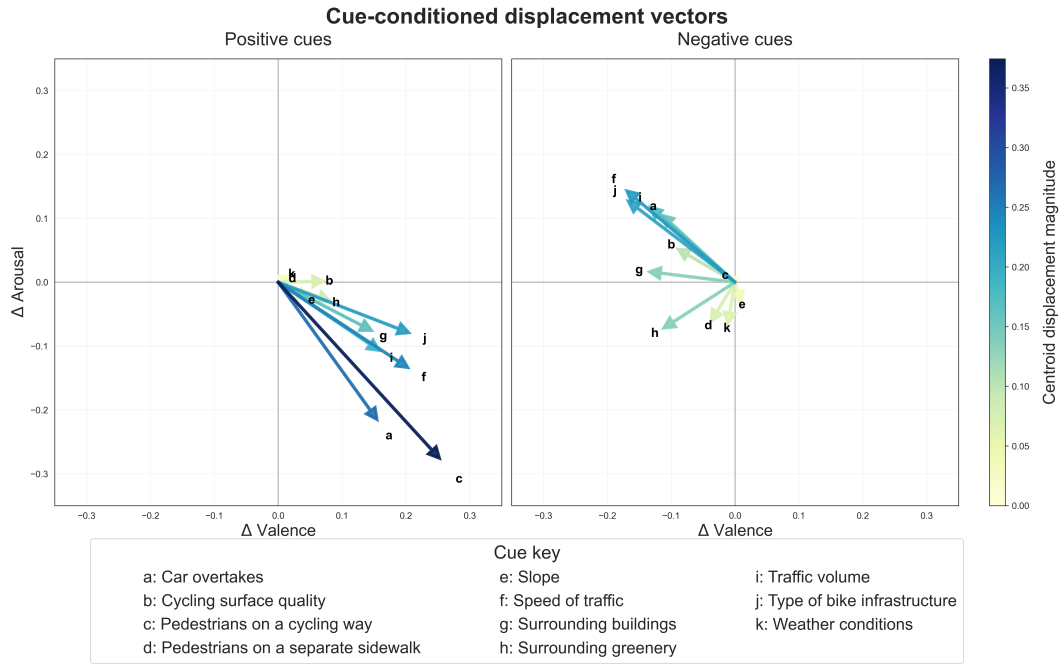


Figure 7 Conditional displacement vectors of tagged environmental factors. Vectors represent the directional shift in valence and arousal associated with specific positive (left) and negative (right) cues relative to the scenario-level mean.

($r = 0.61$), suggesting that scenarios with diverse environmental tags (e.g., conflicting perceptions of "traffic volume" versus "bike infrastructure") are associated with higher affective ambiguity. The conditional displacement vectors (see Figure 7) isolate the directional influence of these factors. Negative factors form a tight cluster pointing toward the high-arousal/negative-valence quadrant. Positive factors shift appraisals toward the relaxation quadrant. "Pedestrians on a cycling way" and "Car overtakes" point toward relaxation (when tagged as positive), suggesting an "absence-of" signal.

Overall, these patterns converge into three recurring structural signatures (see Figure 8). Negative scenarios (e.g., 28, 21) exhibit strong directional structure and low affective ambiguity ($H^Q \approx 0.73 - 0.84$), where high traffic volumes (up to 15,301 AADT) and lack of protected infrastructure align with broad agreement on negative appraisal. Mixed-signal environments (e.g., 6, 12) instead show higher dispersion ($D = 0.74$) and high quadrant entropy ($H^Q \geq 1.32$): stress-mitigating factors (lower speeds, greenery) co-occur with stressors (car overtakes), preventing a single affective consensus. Finally, consensus-positive environments (e.g., 9, 1) shift toward positive valence and lower arousal driven by separated infrastructure and high greenery share ($> 50\%$), but differ in PI profiles (0.85 vs. 1.11).

4.4 Partitioning Sources of Variation (RQ3)

To distinguish scenario-driven signals from observer-driven variability, we stratified the selected metrics by participant subgroups (see Figure 9). The analysis reveals a difference: D, PI, and H^{cue} are systematically sensitive to observer characteristics, whereas AI and H^Q remain largely stable across subgroups. Non-cyclists and infrequent riders show higher values in D, PI, and H^{cue} compared to commuters and frequent riders, which suggests

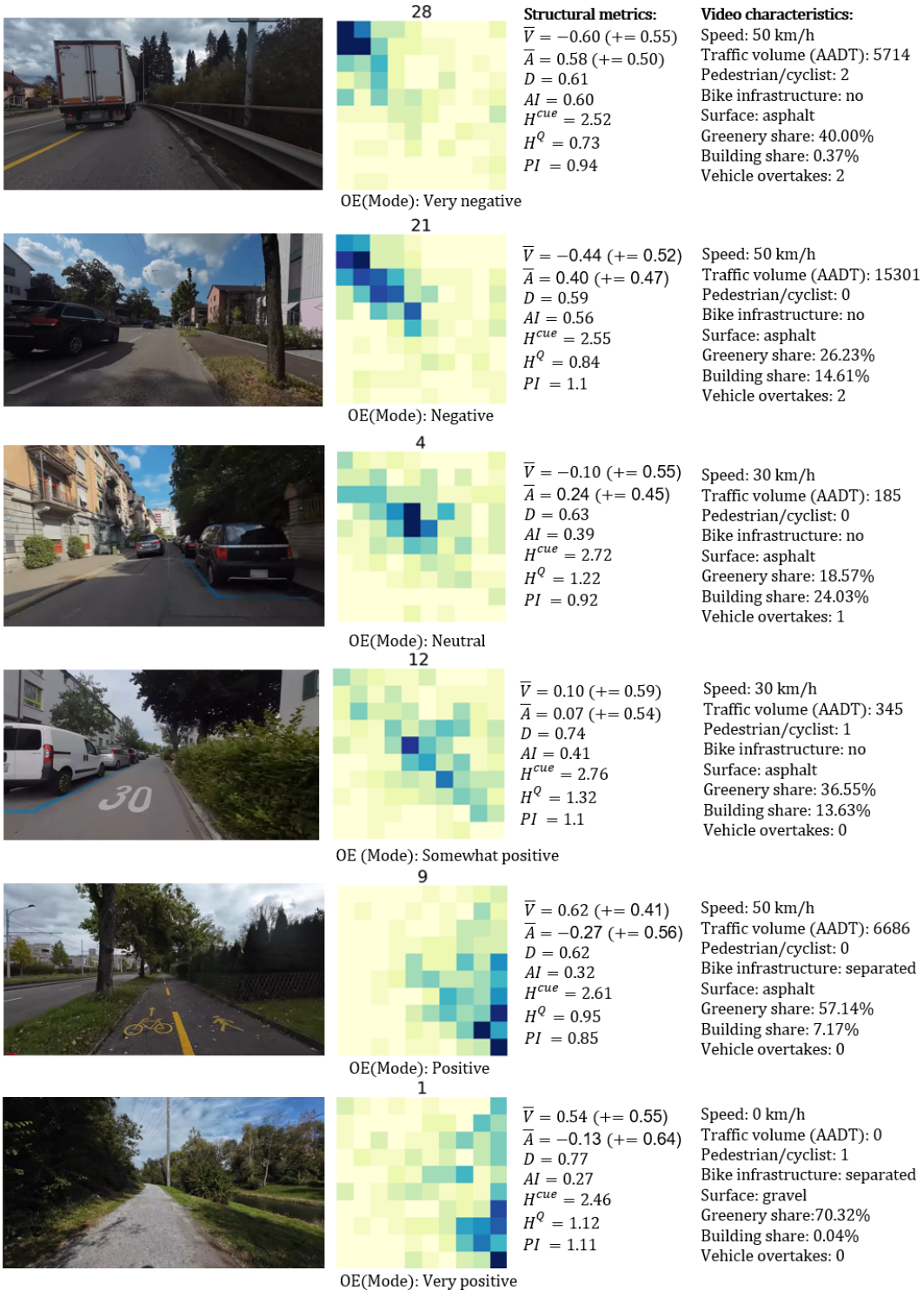


Figure 8 Representative cycling scenarios with varying combinations of physical infrastructure, traffic cues, and geometric properties ($D, AI, PI, H^{cue}, H^{Q}$) that characterize three distinct experience signatures: consensus-positive, structured-stress, and ambiguous-mixed signals.

that experience acts as a tuning mechanism that fosters more shared interpretations. The youngest cohort (18–25) displays significantly higher values across these metrics, consistent with less structured appraisal. Polarization is also lower among rural, recreational, less confident, and female participants. In contrast, the stability of AI and H^Q suggests that directional orientation and affective categorical ambiguity are intrinsic properties of cycling scenarios. In turn, while observer traits modulate the magnitude of variation, the shape of that variation serves as a robust, observer-independent diagnostic of the environmental signal.

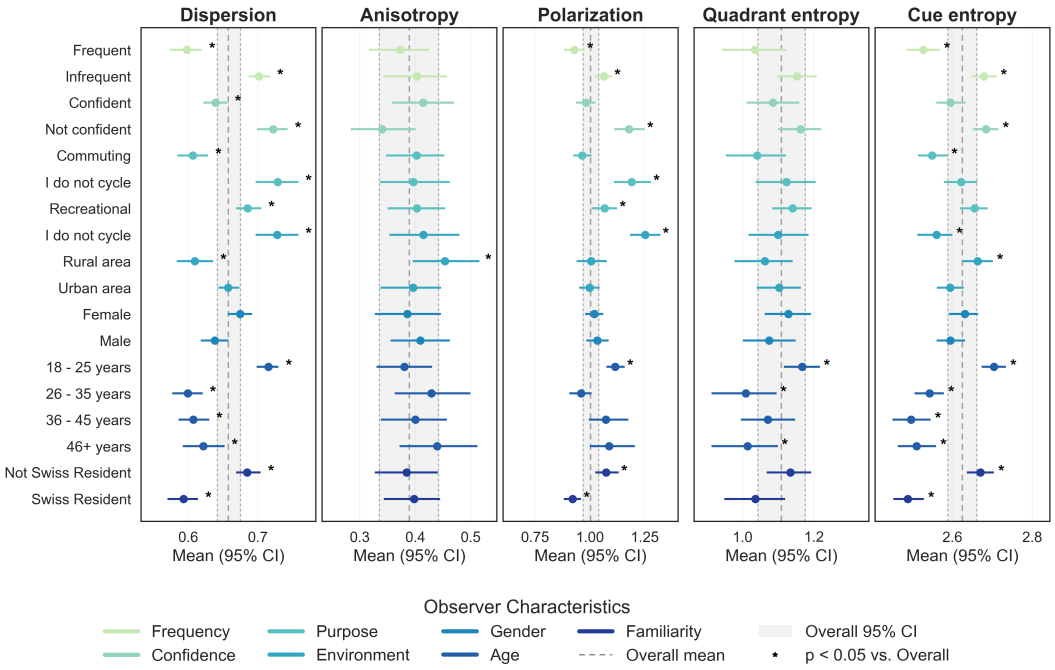


Figure 9 Stability of geometric appraisal metrics across participant subgroups. Individual dots represent subgroup means, while the vertical dashed lines indicate the aggregate sample baseline (mean and 95% CI). Asterisks (*) denote significant deviations from the overall sample mean ($p < 0.05$) based on post-hoc testing.

5 Discussion

This study treats appraisal heterogeneity in cycling experience as a diagnostic signal. We demonstrate that cycling scenarios can be characterized by distinct geometric signatures when projected onto affect space. These signatures provide a novel audit layer for distinguishing between environments that evoke clear stress and those that elicit contested interpretations.

Key Findings. First, we demonstrate that cycling scenarios differ not only in mean affect but in the underlying geometry of appraisal. Our metric set shows that appraisal heterogeneity within the subjective cycling experience is multidimensional with largely independent properties (see Table 3). Second, we identify a structural difference in response heterogeneity. High-stress scenarios typically manifest as clear, directional appraisals (high AI, low D), while positive scenarios result in diffuse appraisals (low AI, high D), appearing as scattered clouds (see Figure 8). Third, we find that affective state entropy correlates with tagged cue

entropy, suggesting an association between signal heterogeneity and competing visual cue interpretations. Fourth, we show a "magnitude vs. shape" dissociation regarding sources of variation. Specifically, while observer experience acts as a tuning mechanism that changes the magnitude of appraisal heterogeneity (often consisting of polarized clusters), the response shape (AI) remains largely robust across participant subgroups (see Figure 9).

Possible Theoretical and Practical Implications. The presented structural response patterns provide complementary evidence for the valence ambiguity model by Mattek et al. [20], who showed across seven datasets of brief affective stimuli (faces, images, words, sentences, sounds; total $N = 195$) that the valence–arousal correlation depends on valence ambiguity, and we replicate this pattern in the cycling domain. More specifically, we found that unambiguous negative stimuli result in valence–arousal coupling (tight linear structure along the stress diagonal). More ambiguous responses to positive stimuli result in decoupled affective dimensions (diffuse, multi-directional variance), with a more heterogeneous response palette (e.g., excitement, relaxation, contentment). For practical intervention design, this difference could mean non-linear improvement dynamics. For example, adding amenities (e.g., greenery) to already "decoupled" (i.e., positive environments) may produce smaller, more scattered shifts as responses diffuse across multiple positive quadrants. On the other hand, removing stressors (e.g., through traffic volume reduction or physical separation) may target the "stress streaks," likely yielding a shift to a more dispersed consensus. This suggests that intervention impact is state-dependent: improving "somewhat positive" infrastructure may not yield linear affective gains, as the underlying geometry shifts from structured to diffuse. This geometric diagnostic enables planners to transition from modeling average satisfaction to auditing environmental legibility, better matching intervention strategies to the underlying structural diversity of cycling experiences. Importantly, such dispersion should not be a reason for concern to planners, as we show it can be attributed to individual differences rather than environmental characteristics.

Contributions. First, we reconceptualized appraisal heterogeneity as an informative signal. We introduced a distributional perspective on cycling experience that preserves the geometric information typically lost in measures of central tendency by separating affective response centroids from their structural geometry (D, AI, PI, H^Q). Second, we demonstrated a dissociation where observer characteristics (e.g., experience) moderate the *magnitude* of variance, yet the *geometric structure* remains largely scenario-driven and robust. Third, we provided a reusable metric set (see Table 1) supported by open-source code that demonstrates how to calculate these structural properties from raw survey data using standard statistical techniques. Finally, our approach complements existing facility typologies and cycling archetypes by enabling a more nuanced evaluation of urban interventions, where a successful street redesign may leave average valence unchanged while significantly reducing affective quadrant or environmental cue entropy, thereby improving the legibility and shared interpretability of the cycling environment.

Limitations. First, the generalizability of our findings is constrained by the sampling strategy. As all clips were recorded in Zurich, the observed signatures reflect the specific infrastructure typology and variability of the local urban form. Broader replication is needed to determine whether these geometric patterns hold in other traffic cultures. Similarly, our results are skewed toward a younger ($72.8\% \leq 35$ y.o.) and more confident (in terms of cycling) population. Given this snowball-recruited sample, we treat the present study as

exploratory and foundational rather than confirmatory. Second, appraisal heterogeneity may be influenced by unmeasured observer traits. Although we accounted for cycling experience and confidence, latent variables such as individual attitudes, risk tolerance, sensation seeking, or prior crash history could further influence appraisal heterogeneity structure. Third, the Affect Grid is bounded by design, which constrains variance at the affective extremes. Its standard quadrant labels (e.g., "stress," "boredom") were also developed for general affect research and may not perfectly capture the specific sensations of cycling, though using validated terminology keeps our metrics comparable across studies. We also treat the absence of a dominance dimension (sampled by SAM but not by the Affect Grid) as a deliberate scope choice, traded against respondent fatigue across our 30 scenarios. Fourth, while first-person video captures dynamic visual cues, it lacks the vestibular feedback, physical effort, and personal control and risk that shape in-situ arousal, which was a deliberate trade-off to favor a broader outreach over full physical immersion. This, combined with the exclusion of prolonged static traffic congestion from our stimuli, likely explains the absence of the "boredom" quadrant (low valence, low arousal) in our dataset. Translating the framework to in-situ or immersive VR settings is therefore a necessary next step. The physical effort of real or active-VR cycling could elevate baseline arousal and shift or compress the geometric signatures we observe, leaving it unclear whether our structural patterns persist under fully embodied conditions.

6 Conclusions and Outlook

This paper introduces a geometric framework for analyzing subjective cycling experiences, moving beyond central tendency to quantify the structure of appraisal heterogeneity. The subset of metrics utilized proved largely robust across observer subgroups, suggesting they reflect intrinsic scenario properties rather than subgroup preference. We identified three distinct geometric signatures: diffuse consensus (positive environments), linear structured stress (negative environments), and polarized or ambiguous appraisals (mixed-signal environments). This lets practitioners distinguish between environments that are consistently stressful and those that potentially elicit multiple interpretations.

Future research should extend this geometric framework to in-situ experiments and diverse traffic cultures to test the universality of these structural signatures. Confirmatory follow-up studies should also employ a priori power and sample-size calculations and apply confirmatory statistical models (e.g., mixed-effects regressions) to formally test which environmental and observer variables predict the geometric signatures identified here. While our video-based methodology effectively captures dynamic visual-cognitive cycling experience appraisals, incorporating the physiological effort and vestibular feedback of real-world cycling or developing a cycling-relevant affect grid terminology remain the next research steps. Another direction is adding a dominance dimension to the appraisal. It could help differentiate our "structured-stress" signature into active stress (anger, frustration; high dominance) and passive stress (fear, helplessness; low dominance), though whether the resulting gain in resolution outweighs the added respondent burden in repeated-measures designs remains an open question. Ultimately, moving from a model of average satisfaction to a model of environmental legibility will support the design of cycling environments that are both inclusive and affectively coherent.

References

1 Frauke Luise Berghoefer and Mark Vollrath. Prefer what you like? Evaluation and preference of cycling infrastructures in a bicycle simulator. *Journal of Safety Research*, 87:157–167, 2023. doi:10.1016/j.jsr.2023.09.013.

2 Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. doi:10.1016/0005-7916(94)90063-9.

3 Laura Cabral and Amy M. Kim. An empirical reappraisal of the four types of cyclists. *Transportation Research Part A: Policy and Practice*, 137:206–221, 2020. doi:10.1016/j.tra.2020.05.006.

4 Ana Karina Christ, Miguel Costa, Manuel Marques, Carlos Roque, and Filipe Moura. Perceiving objective cycling safety: a systematic literature review. *Transportation Research Procedia*, 72:1380–1387, 2023. doi:10.1016/j.trpro.2023.11.601.

5 Miguel Costa, Manuel Marques, Felix Wilhelm Siebert, Carlos Lima Azevedo, and Filipe Moura. Bridging the gap: a scoping review exploring the relationship between objective and subjective cycling safety. *Transport Reviews*, 45(6):993–1015, 2025. doi:10.1080/01441647.2025.2537200.

6 Tineke De Jong and Aslak Fyhri. Spatial characteristics of unpleasant cycling experiences. *Journal of Transport Geography*, 112:103646, 2023. doi:10.1016/j.jtrangeo.2023.103646.

7 Jennifer Dill and Nathan McNeil. Revisiting the Four Types of Cyclists: Findings from a National Survey. *Transportation Research Record: Journal of the Transportation Research Board*, 2587(1):90–99, 2016. doi:10.3141/2587-11.

8 Haining Ding, Chenxi Wang, and Michal Gath-Morad. How built environment shapes cycling experience: A multi-scale review in historical urban contexts, 2025. arXiv:2509.15657. doi:10.48550/arXiv.2509.15657.

9 J. J. Gibson. The theory of Affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

10 Ayda Grisiute, Nina Wiedemann, Pieter Herthogs, and Martin Raubal. An ontology-based approach for harmonizing metrics in bike network evaluations. *Computers, Environment and Urban Systems*, 113:102178, 2024. doi:10.1016/j.compenvurbsys.2024.102178.

11 Eva Heinen, Bert Van Wee, and Kees Maat. Commuting by Bicycle: An Overview of the Literature. *Transport Reviews*, 30(1):59–96, 2010. doi:10.1080/01441640903187001.

12 Anja Katharina Huemer, Luzie Marianne Rosenboom, Melina Naujoks, and Elise Banach. Testing cycling infrastructure layout in virtual environments: An examination from a bicycle rider's perspective in simulation and online. *Transportation Research Interdisciplinary Perspectives*, 14:100586, 2022. doi:10.1016/j.trip.2022.100586.

13 Anjani Kalra, Tommy Lim, Lauren Pearson, and Ben Beck. Methods used to capture subjective user experiences in adults while riding bicycles: a scoping review. *Transport Reviews*, 43(3):453–477, 2023. doi:10.1080/01441647.2022.2123064.

14 Fatima Kchour, Salvatore Cafiso, and Giuseppina Pappalardo. Understanding Cyclists' Visual Behavior Using Eye-Tracking Technology: A Systematic Review. *Sensors*, 25(1):22, 2024. doi:10.3390/s25010022.

15 Peter Kuppens, Francis Tuerlinckx, Michelle Yik, Peter Koval, Joachim Coosemans, Kevin J. Zeng, and James A. Russell. The Relation Between Valence and Arousal in Subjective Experience Varies With Personality and Culture. *Journal of Personality*, 85(4):530–542, 2017. doi:10.1111/jopy.12258.

16 Tommy Lim, Anjani Kalra, Jason Thompson, Joanne Caldwell Odgers, and Ben Beck. Physiological measures of bicyclists' subjective experiences: A scoping review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 90:365–381, 2022. doi:10.1016/j.trf.2022.09.007.

17 Tommy Lim, Lauren Pearson, Jason Thompson, Joanne Caldwell Odgers, and Ben Beck. Establishing normative scores of affect, safety, and comfort for bicycling infrastructure and traffic conditions. *Transportation Research Part F: Traffic Psychology and Behaviour*, 114:337–353, 2025. doi:10.1016/j.trf.2025.06.006.

18 Tommy Lim, Jason Thompson, Lauren Pearson, Joanne Caldwell Odgers, and Ben Beck. Effects of within-trip subjective experiences on travel satisfaction and travel mode choice: A conceptual framework. *Transportation Research Part F: Traffic Psychology and Behaviour*, 104:201–216, 2024. doi:10.1016/j.trf.2024.05.024.

19 Kevin Lynch. *The Image of the City*. Publications of the Joint Center for Urban Studies. MIT Press, Cambridge, MA, 1960.

20 Alison M. Mattek, George L. Wolford, and Paul J. Whalen. A Mathematical Model Captures the Structure of Subjective Affect. *Perspectives on Psychological Science*, 12(3):508–526, 2017. doi:10.1177/1745691616685863.

21 Andrii Matviienko, Hajris Hoxha, and Max Mühlhäuser. What does it mean to cycle in Virtual Reality? Exploring cycling fidelity and control of VR bicycle simulators. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3544548.3581050.

22 Jun Qi, Xueqiong Tang, and Yongju Luo. Affordances of scenic cycleways: How recreational cyclists interact with different environments. *Urban Forestry & Urban Greening*, 64:127270, 2021. doi:10.1016/j.ufug.2021.127270.

23 Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M. Parkhi. EgoBlur: Responsible innovation in Aria, 2023. arXiv:2308.13093. doi:10.48550/arXiv.2308.13093.

24 Martin Raubal. Wayfinding: Affordances and Agent Simulation. In Shashi Shekhar and Hui Xiong, editors, *Encyclopedia of GIS*, pages 1243–1246. Springer US, Boston, MA, 2008. doi:10.1007/978-0-387-35973-1_1469.

25 James Russell, Anna Weiss, and G. Mendelsohn. Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57:493–502, 1989. doi:10.1037/0022-3514.57.3.493.

26 Sonja Schneider and Klaus Bengler. Evaluating behavioral validity in traffic simulators. In *Proceedings of TRA2020, the 8th Transport Research Arena*, Helsinki, Finland, 2020.

27 Justin Spinney. Cycling the City: Movement, Meaning and Method. *Geography Compass*, 3(2):817–835, 2009. doi:10.1111/j.1749-8198.2008.00211.x.

28 Markus Tomzig, Johanna Wörle, Alexandra Neukum, and Martin Baumann. The arousal paradox in critical task performance in automated driving during sleep inertia using a quasi experimental approach. *Scientific Reports*, 15(1):20338, 2025. doi:10.1038/s41598-025-08726-4.

29 Ran Zhang, Marco Te Brömmelstroet, Anna Nikolaeva, and George Liu. Cycling subjective experience: A conceptual framework and methods review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 101:142–159, 2024. doi:10.1016/j.trf.2023.12.021.

Understanding Building Layouts from Minimal 360° Views: Effects of Viewpoint Overlap, Sequencing, and Landmarks

Ahmed Aly
Institute for Geoinformatics, Universität Münster, Germany

Jakub Krukar
Institute for Geoinformatics, Universität Münster, Germany

Abstract

While 360° panoramic virtual tours are widely used for remote spatial exploration, there is no standardized method for determining the placement or sequencing of such viewpoints. Computational solutions like the Art Gallery Problem can determine minimal viewpoints that ensure the building's coverage, however cognitive research has not addressed whether these are sufficient to support the viewer's understanding of the space. In other words, would the hypothetical guards understand their gallery's layout? Drawing on the Network of Reference Frames theory and prior research on spatial integration mechanisms and associated integration costs, this study investigates how three design factors – visual overlap, route-based sequencing, and landmark availability – influence layout understanding from a minimal set of views. A reproducible web-based experimental framework was developed to enable systematic manipulation of viewpoint configurations and detailed behavioral logging. Fifty participants explored five systematically varied configurations across five building layouts and were tasked with recognizing the correct layout among visually similar alternatives, reporting confidence after each decision. Monetary incentives were used to encourage attentive exploration and constrain strategy use. Results revealed that while individual spatial ability is a strong predictor of layout understanding ($p < .001$), viewpoint design significantly modulates that understanding. Specifically, visual overlap between consecutive views significantly enhanced layout understanding only when paired with additional cues – namely landmarks or route-based sequencing. Among these, landmarks significantly improved layout recognition ($p < .05$), unlike route-based sequencing. Moreover, combining both cues yielded no additional benefit. These findings demonstrate how viewpoint configurations can be designed to better support layout understanding from a minimal set of views. In addition, the proposed replicable, web-based experimental pipeline enables systematic evaluation and optimization of viewpoint designs used in spatial learning and navigation tools.

2012 ACM Subject Classification Human-centered computing → Empirical studies in HCI; Human-centered computing → Virtual reality; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Virtual tours, Art Gallery Problem, museum guard problem, isovists, layout understanding, 360° panoramas

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.9

Supplementary Material *Dataset (Raw data)*: `https://osf.io/8e7pf/`
Software (Experiment code): `https://github.com/ahmedsaly/LayoutRecStudy`
archived at `swh:1:dir:855b206332b133e5573f2a51680b197ee7850902`

Funding This project was supported through the JICE Seed Funding by the Joint Institute of Individualisation in a Changing Environment.

© Ahmed Aly and Jakub Krukar;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 9; pp. 9:1–9:23

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Virtual representations of architectural spaces have been increasingly used across multiple domains, including real estate, architecture, and interior design [43]. A particularly widespread approach is the use of *virtual tours* – interactive systems composed of sequential images, often in 360° panoramic format, that enable users to explore built environments remotely. The popularity of these systems, supported by commercial software such as Matterport or 3DVista, lies in their utility: users can form an impression of spatial structure and layout without physically entering the space.

In research contexts, similar image-based paradigms are used to study spatial cognition, navigation, and environmental understanding. For example, researchers simulate building exploration by presenting participants with a set of images [46, 24]. However, despite the use of such paradigms, there remains no standardized method for determining *where* to place viewpoint images, *how many* are required, or *what sequence* best supports spatial understanding. This highlights a central research gap: *How can viewpoint systems be systematically selected and organized to support spatial layout understanding?*

Addressing this question has both theoretical and practical relevance. From a theoretical perspective, it contributes to modeling how humans acquire survey knowledge from fragmented visual input – concepts central to theories describing how we integrate information across reference frames (see Section 2.1). Practically, it can inform the design of spatial learning tools, legible building layouts, indoor navigation systems, and architectural visualizations based on panoramic imagery. The primary contribution of this work is an empirical investigation of which viewpoint-system design factors support layout understanding from a minimal set of views.

An additional methodological contribution is a reproducible, web-based experimental framework for panoramic layout-understanding studies. It enables systematic manipulation of viewpoint configurations while capturing detailed behavioral responses of participants, and incorporates a monetary incentive structure to encourage attentive exploration. Using this framework, viewpoint positions were optimized using the Art Gallery Problem (AGP) algorithm to calculate the minimum number of views required for full geometric coverage (see Section 2.3). Across these optimized positions, the degree of overlap, the sequencing strategy, and the availability of landmarks were systematically added to test their impact.

Our findings reveal that layout understanding was strongly predicted by individual spatial ability and was significantly improved when landmarks were available within overlapping views. Furthermore, visual overlap alone was found to be insufficient without additional cues, and the monetary incentive structure also influenced exploration strategies.

2 Background

2.1 Reference frames and spatial integration

To understand how individuals integrate static panoramic images into a coherent mental model, we draw on the *Network of Reference Frames* theory [33]. It proposes that spatial knowledge is not stored in a single global coordinate system, but rather as a network of locally encoded *vista spaces* – areas visible from a single viewpoint [33, 37]. Each vista space is stored in its own reference frame, and transitions between vista spaces are represented by *perspective shifts* involving translation and rotation required to move from one reference frame to another. Comprehension of an environment, therefore, depends on the successful formation of these shifts [33].

Forming survey knowledge from panoramic imagery thus fundamentally relies on *spatial integration* – the mental unification of separately experienced vista spaces into a coherent environmental representation [33, 16, 22]. This integration relies on working-memory demands during mental transformation and alignment, introducing what is known as *integration cost* [34]. Integration costs are reflected in reduced spatial accuracy, increased task completion time, or higher error rates during recall [34]. These costs depend on how easily viewpoints can be aligned and linked into a consistent representation, and can be reduced by cues that support perspective-shift formation and reference-frame alignment.

For adjacent vista spaces, *visual overlap* provides a primary mechanism forming perspective shifts: when two vista spaces share a direct visual connection, the shared geometry provides the necessary translational and rotational components to link the frames [33]. As integration costs diminish when alignment occurs *during encoding* rather than being postponed to retrieval [34], integration during encoding is operationalized through an *overlap-based sequence*: presenting overlapping views consecutively so that each new panorama shares visual information with the previous one. In our exploratory design, this overlap-based sequence serves as the baseline for testing whether additional cues further facilitate integration and improve layout understanding.

2.2 Cues beyond overlap: landmarks and route-based sequencing

Beyond overlap, additional cues can support integration across views and help build a coherent layout representation. One such cue is *global landmarks*: distinctive objects visible from multiple locations can act as common anchors. Self-localization relative to a landmark allows the observer to reorient across different vista spaces, effectively providing the perspective shift between reference frames [33]. Landmarks within overlapping regions provide stable reference points that help align egocentric views to a shared allocentric structure, improving memory precision [1, 15, 25] and thus reducing integration costs [34]. In the context of panoramic viewpoint systems, this motivates placing landmarks so they remain visible across consecutive, overlapping views, providing persistent anchors across view transitions.

Another cue is continuous updating during movement. When moving between non-adjacent vista spaces, individuals can track their position and orientation relative to the last encoded vista via path integration, accumulating translational and rotational information to achieve the necessary perspective shift [32, 33]. In panoramic viewpoint systems, transitions between views are inherently discontinuous and therefore resemble teleportation-like navigation. Consistent with VR locomotion research, comparisons of teleportation with continuous movement indicate that discontinuous transitions can impair spatial updating and orientation and can bias spatial judgments such as distance perception [6, 29]. Accordingly, many studies focus on improving how teleportation is executed – for example, adding orientation-preserving mechanisms, rotational self-motion cues, or visual transition/flow cues to reduce disorientation and better support spatial updating [31, 42, 3].

However, these advances mainly refine transition mechanics rather than the placement of teleport targets. Recent work on scene-aware teleport graphs proposes structured sets of teleport nodes derived from environmental geometry [28], but how to select a minimal set of viewpoints that optimally supports users' understanding of the global layout remains largely unexplored. Our focus here is not on rendering smooth transitions between viewpoints, but on how a minimal set of discrete viewpoints is selected and ordered to support layout understanding. At the same time, continuous visual flow between viewpoints, even in the absence of active bodily movement, appears to be a key driver of spatial learning: smooth transitions between views enhance small-scale spatial learning, with passive viewing of these

transitions performing as well as active movement [20]. A straightforward way to approximate some of this continuity in panoramic viewpoint systems is to structure the sequence of views so that it follows a plausible route through the environment. In this study, a *route-based sequence* was therefore tested: the same set of overlapping views was ordered to mirror how a person might naturally explore the building.

Finally, integration efficiency is subject to individual differences. The ability to form and use perspective shifts varies by individual spatial ability [18]. Participants with stronger spatial abilities, such as higher mental rotation skills or a better sense of direction, demonstrate quicker and more accurate integration [18]. Therefore, spatial ability is measured and accounted for when analyzing differences in layout understanding between participants.

2.3 Methodological foundations: minimal viewpoints and incentives

One methodological concern is limiting the number of viewpoints while still supporting reliable layout understanding. The cognitive demands of spatial integration increase with the number of distinct vista spaces that must be linked into a coherent representation [33]. For example, pointing errors increase with the number of vista spaces traversed [49]. This motivates a key design goal for panoramic viewpoint systems: fewer, well-selected viewpoints should reduce cognitive load and support more efficient layout understanding. The challenge, however, is how to identify the smallest set of viewpoints that is sufficient for understanding the space. In the absence of a systematic method for determining what people need to see to understand a space, our starting assumption is to require *full visual coverage* of the space. This does not assume that full geometric coverage equals cognitive understanding; rather, it serves as a principled baseline for testing which further design factors support understanding. Under this condition, the question becomes: how can we identify the smallest set of viewpoints that collectively provides full coverage? This is a computational optimization problem, and it naturally maps to the Art Gallery Problem (AGP) [8, 39].

The AGP addresses the question of how many "guards" with 360° visibility are needed to observe every point in a polygonal environment of an art gallery [8, 39]. While classical theory gives general bounds on how many guards are needed, computing an *optimal* solution is NP-hard, especially when guards are allowed to be placed anywhere inside the polygon (point guards) [39]. A practical and widely used approach is the guard–witness formulation, which reduces coverage to a finite set of candidate guards and witness points and solves the resulting set-cover instance via Integer Linear Programming (ILP), iteratively refining the witness set until full coverage is achieved. Algorithm 966 [50] provides a robust implementation of this strategy and has been shown to solve thousands of benchmark instances efficiently on standard hardware, achieving optimal solutions in over 95% of more than 2,700 tested cases (including polygons with up to 2,500 vertices) [50]. In this study, this method is used to compute the minimum set of viewpoints required for full spatial coverage.

A second methodological concern is maintaining participant attention under potentially high cognitive demands. Because spatial integration is effortful and time-consuming, participants may adopt different viewing strategies depending on task framing. Monetary incentives can increase motivation and attentional effort, especially in tasks that require the deliberate allocation of cognitive resources [4, 21]. Therefore, in this study, a monetary incentive structure is used to encourage attentive exploration and integration across the available views rather than superficial task completion.

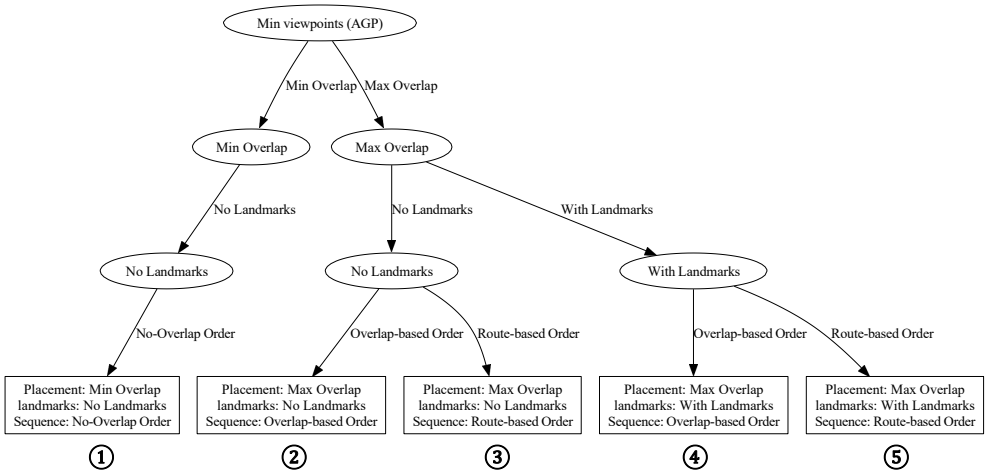


Figure 1 Overview of the five experimental conditions, varying overlap, sequencing strategy, and landmark presence.

3 Methods

3.1 Research question and experimental conditions

Building on the integration mechanisms reviewed in Section 2, to examine how viewpoint-system design influences *layout understanding* when indoor environments are explored from a minimal set of 360° panoramas, the following research question is posed:

(RQ). *How do viewpoint overlap, sequencing strategy, and landmark cues affect layout understanding from a minimal set of panoramic views?*

Here, *layout understanding* refers to the ability to form a mental representation of a building's overall structure (the number and arrangement of rooms or spaces, and their relative positions), sufficient to recognize a layout among visually similar alternatives.

To address this question, the following hypotheses are proposed:

- **H1 (Overlap).** Visual overlap between consecutive views is essential for layout understanding; lower layout understanding is expected when consecutive views provide no overlap compared to when overlap is high.
- **H2 (Cues beyond overlap).** When overlap is available, additional cues further improve layout understanding:
  - **H2a (Route-based order).** A route-based sequence is expected to yield higher layout understanding than an overlap-based sequence.
  - **H2b (Landmarks).** Landmark availability within overlap regions is expected to yield higher layout understanding than an overlap-based sequence without landmarks.
  - **H2c (Combined cues).** Combining route-based order and landmarks is expected to yield the strongest layout understanding.
- **H3 (Individual differences).** Higher individual spatial ability predicts better layout understanding.

To test these hypotheses, five experimental conditions were created and implemented across five building layouts. The conditions differed along three factors: viewpoint placement, viewing sequence, and the presence of landmarks. Figure 1 summarizes the construction

Table 1 Overlap statistics for the min- and max-overlap placements. *Total overlap* sums all pairwise isovist overlaps as a percentage of building area; regions seen by several viewpoints are counted once per pair, so the value can exceed 100%. *Overlapping viewpoints* is the number of viewpoints sharing overlap with at least one other, out of the total.

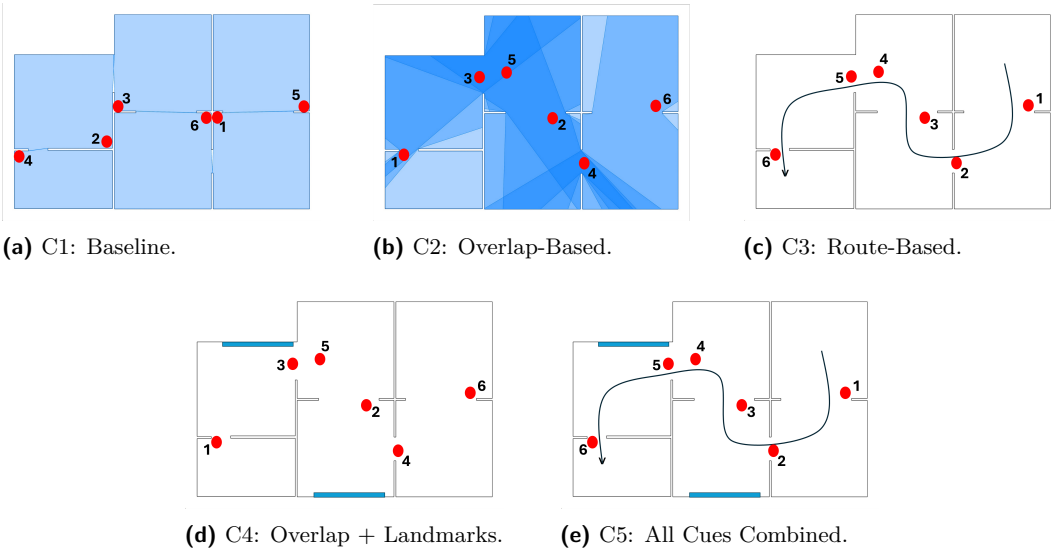
	Total overlap		Overlapping viewpoints	
Building	**Min**	**Max**	**Min**	**Max**
B1	3.18%	182.78%	2/4	4/4
B2	1.43%	285.93%	4/6	6/6
B3	3.45%	209.90%	2/5	5/5
B4	29.11%	92.85%	3/4	4/4
B5	0.70%	60.23%	2/6	6/6

pipeline and the resulting five conditions. The *minimal viewpoint number* required for full visual coverage was computed via an AGP solution (Section 2.3) and was held constant across conditions (within each building layout). These factors were then manipulated under the same full-coverage constraint:

- **Overlap level (viewpoint placement).** A *min-overlap* placement was constructed in which visual overlap of area seen between any pair of viewpoints was minimized as far as the layout permitted. A *max-overlap* placement was constructed in which overlap was maximized, aiming to enlarge the regions that can be jointly seen from multiple viewpoints and to increase the number of viewpoints that share overlap. Because strict "no overlap" between all viewpoint pairs was not always achievable, *min overlap* denotes the lowest feasible overall overlap, whereas *max overlap* denotes a placement in which overlap was made as extensive as feasible. The contrast between placements is summarized in Table 1 and visualized via isovists in Figure 2(a–b).
- **Sequencing strategy (transitions).** In an *overlap-based sequence*, views were ordered so that consecutive views overlapped. In a *route-based sequence*, consecutive overlap was preserved while the order was adjusted to approximate a plausible walking path through the environment. The walking path was designed manually, covering all rooms in the layout, visiting each only once, and staying in the center of the spaces (following findings that natural movement is biased towards *lowest isovist drift* [9].)
- **Landmark availability.** Distinctive landmarks were added and were placed to remain visible across consecutive overlapping views (i.e., within overlap regions).

Experimental conditions. Figure 2 illustrates the manipulations for one example layout. Each condition was implemented on each building layout as follows:

- **C1: Min overlap + no-overlap order + no landmarks (baseline).** A min-overlap placement is used, and views are ordered so that successive views do not overlap and that the sequence does not approximate a route.
- **C2: Max overlap + overlap-based order + no landmarks.** A max-overlap placement is used, and views are ordered so that each view overlaps with the next.
- **C3: Max overlap + route-based order + no landmarks.** The same placement as C2 is used, but views are ordered to approximate a route-like path while preserving consecutive overlap.
- **C4: Max overlap + overlap-based order + landmarks.** The same placement and order as C2 are used, but landmarks are added to be visible across overlap regions.
- **C5: Max overlap + route-based order + landmarks.** The same placement as C2 is used, with route-like ordering and landmarks, while maintaining consecutive overlap.



(a) C1: Baseline. **(b)** C2: Overlap-Based. **(c)** C3: Route-Based.

(d) C4: Overlap + Landmarks. **(e)** C5: All Cues Combined.

Figure 2 Five experimental conditions for Building B2. Red dots show viewpoints (numbers = order); blue bars = landmarks. In (a)–(b), semi-transparent isovists (not shown to participants) quantify overlap: (a) min-overlap placement, (b) max-overlap. C2–C5 use max-overlap placements; C1 uses min-overlap.

These conditions allow systematic testing of the proposed hypotheses: H1 is tested by comparing C1 against the overlap-rich conditions (C2 to C5); lower layout understanding in C1 would confirm the essential role of overlap in layout understanding. H2a is tested by comparing C2 vs. C3, where an improvement in C3 would indicate that route-based sequencing provides additional benefits beyond overlap-based sequence alone. H2b is tested by comparing C2 vs. C4, where an improvement in C4 would suggest that landmarks enhance layout understanding beyond overlap alone. H2c is tested by comparing C5 against C3 and C4, where an improvement in C5 would indicate additive or synergistic benefits from combining route-based sequencing and landmark cues. Spatial ability (H3) is included as an individual-differences predictor in the analysis.

3.2 Participants

Participants were recruited via Prolific and were required to be located in Germany or the Netherlands, fluent in English, and to have a Prolific approval rate between 95% and 100%; the study was restricted to desktop devices. Fifty-four participants were recruited, and $n = 50$ were retained after predefined attention and comprehension checks (34 male, 16 female; age 18–55, $M = 30.5$, $SD = 9.0$). The full set of exclusion criteria is provided in Appendix A.

3.3 Experimental platform

A browser-based framework was developed to support controlled presentation of panoramic views, enforcement of task constraints, and fine-grained behavioral logging in an online study. The framework was developed using `jsPsych` for experimental logic and flow [10] and

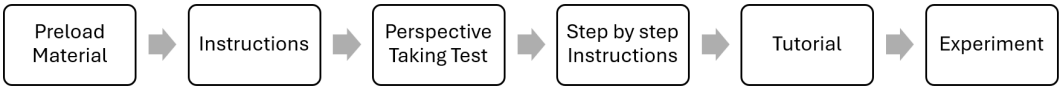


Figure 3 Overview of the six-phase experimental flow implemented in the study.

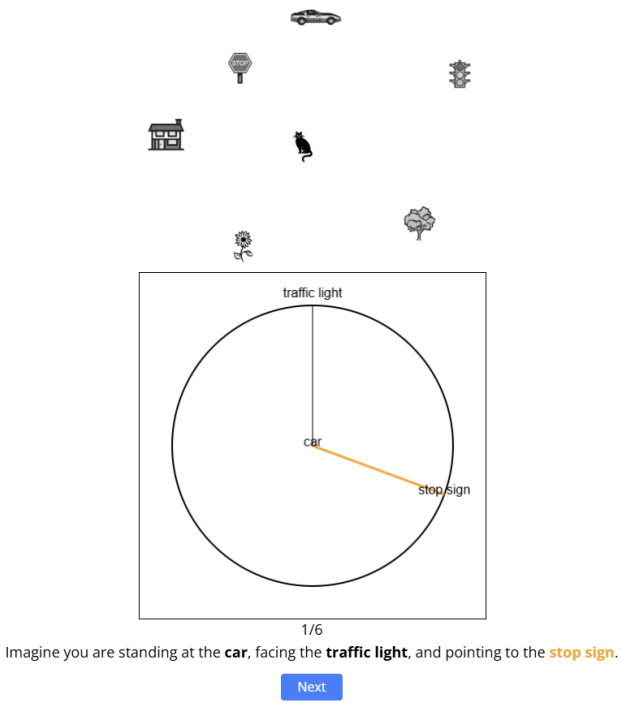


Figure 4 Perspective-Taking Test (PTT) trial. Participants imagined standing at one object (e.g., car), facing a second (e.g., traffic light), and indicated the direction to a third (e.g., stop sign).

`Panolens.js`[1] for rendering panoramic images. Interaction data were logged in real time using `DataPipe` [11], including view transitions, timestamps, layout selections, confidence ratings, and viewport behavior (active exploration vs. static viewing, explored view segments, and segment-level viewing durations). Data were stored on the Open Science Framework (OSF). The front end was deployed via GitHub Pages, and participants were recruited through Prolific.

3.4 Procedure and experimental flow

The study followed a six-phase procedure (Figure 3). First, all stimuli and interface assets were preloaded on participants' browser to avoid latency-related differences in loading and interaction timing. Second, general instructions were presented for both the spatial-ability assessment and the layout-recognition task. Third, spatial ability was assessed using a computerized Perspective-Taking Test (PTT; Figure 4). The PTT consisted of six trials in which participants were asked to imagine standing at one object, facing a second, and indicating the direction to a third object by clicking on a circular response display. Angular

[1] `https://github.com/pchen66/panolens.js`

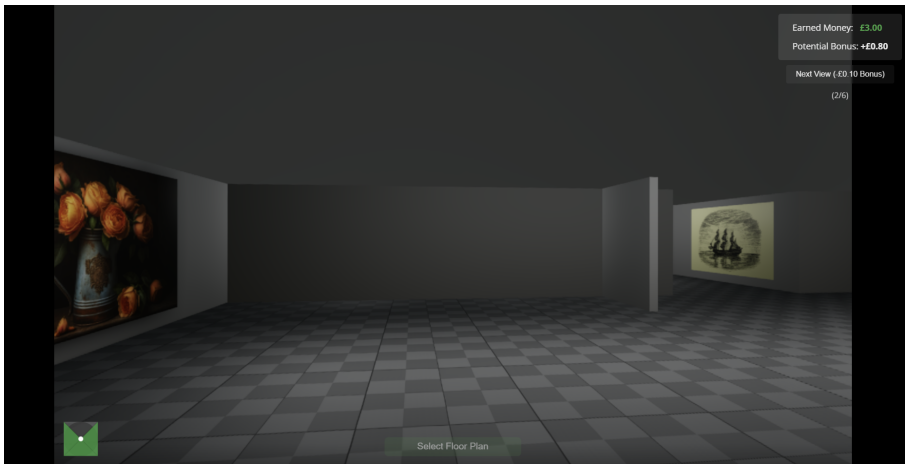


Figure 5 Main experimental interface displaying the 360° panorama, progress indicator (bottom-left), bonus status (top-right), and layout selection button (bottom-center).

error was computed per trial and aggregated into a participant-level score, which was normalized as a z-score for between-participant comparisons (lower values indicating better spatial ability). The PTT was selected as it measures rotational perspective-taking, which is critical for integrating spatial information across multiple viewpoints. The original Spatial Orientation Test (SOT) was introduced in prior work [26, 17], and the computerized version used here followed a later adaptation [12].

Fourth, step-by-step task instructions were provided for the panoramic interface, exploration mechanics, reward logic, and layout selection process. Fifth, a practice tutorial trial was completed to familiarize participants with the full sequence of exploration, decision-making, and confidence reporting. The tutorial used a simplified two-view layout with two alternative floor plans to choose between, where the two alternatives differed in a single room. Sixth, the main task was completed across five trials, each involving one building under one experimental condition. On each trial, 360° panoramic views were explored (Figure 5) before a layout was selected from a set of alternatives, after which confidence was rated on a five-point scale (Figure 6) and trial feedback (correct/incorrect, correct answer, and bonus balance) was presented. Layout understanding was operationalized via correct layout recognition among visually similar alternatives.

During the main task, a standardized panoramic interface was used. Each panorama was subdivided into four 90° segments, and all segments (i.e., the full scene) were required to be viewed at least briefly before participants could advance to the next panorama or select a layout. Viewing progress was displayed via an indicator. Panoramas were explored via mouse dragging or keyboard input. The interface was designed to maintain visual consistency across screen sizes and browser types.

3.5 Stimuli, layouts, and layout alternatives

Five building layouts were selected from the HouseExpo dataset [30]. As the raw layouts contained geometric noise from SLAM-based scanning artifacts, vertex simplification and wall regularization were applied using a custom script to improve architectural realism and to ensure compatibility with AGP-based viewpoint computation. Candidate layouts were screened to support all five experimental conditions described in Section 3.1. Selection was

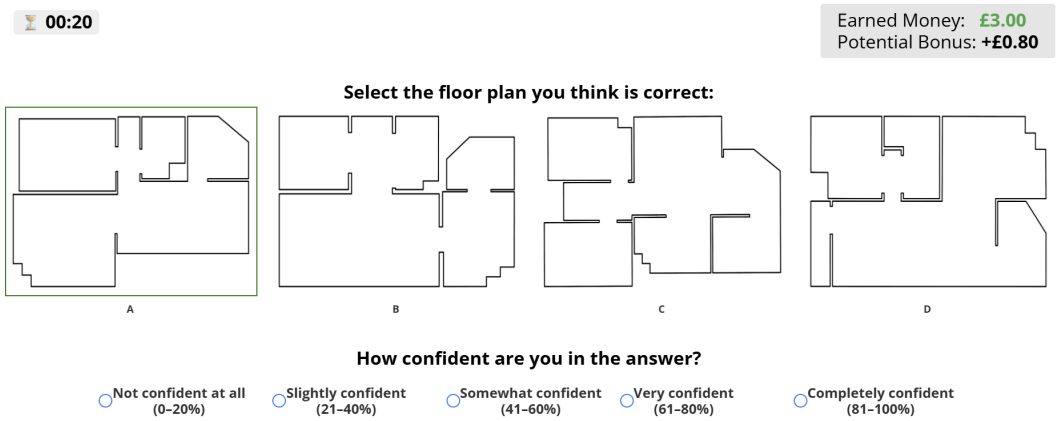


Figure 6 Layout-choice screen (example Building B4). Participants selected the correct layout among three visually similar alternatives within a 40 s decision window, then reported confidence on a five-point scale.

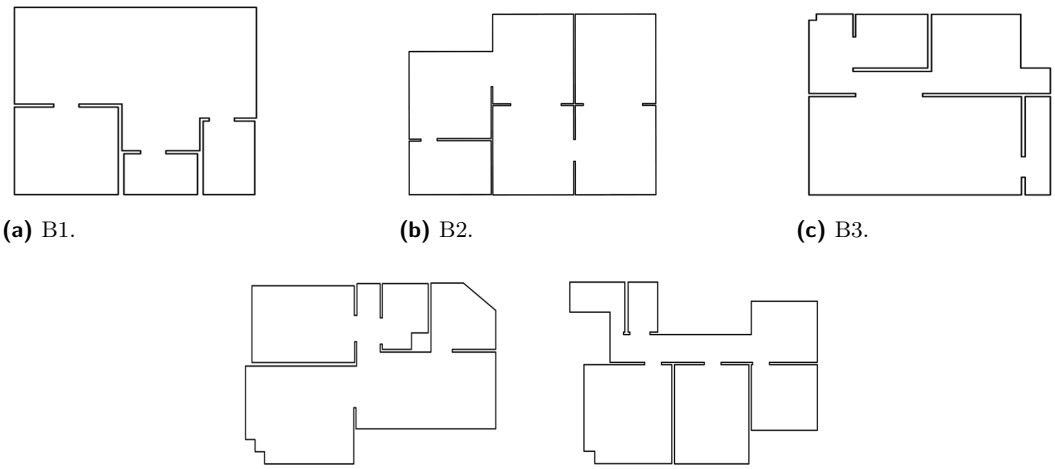
Table 2 Condition assignments across layouts for the five experimental combinations. The combinations forming a balanced Latin square in which each condition appears exactly once per building and once per combination.

	Building				
Combination	**B1**	**B2**	**B3**	**B4**	**B5**
1	C2	C4	C1	C3	C5
2	C5	C1	C3	C2	C4
3	C3	C5	C4	C1	C2
4	C1	C2	C5	C4	C3
5	C4	C3	C2	C5	C1

further constrained such that the AGP 966 algorithm yielded between 4 and 6 viewpoints per layout. The final five layouts were selected to be comparable in overall size yet structurally distinct. Selected rooms were reshaped manually to introduce distinctive geometric features (e.g., diagonal corners, L-shapes, or elongated rectangles), resulting in layouts that vary in the number of distinctive spaces (with some where most rooms were unique, and some with only one or two highly distinctive rooms). Layouts were also inspected to ensure that no two viewpoints within a layout were redundant or visually ambiguous. The final set of five layouts is shown in Figure 7.

To counterbalance the five buildings and five experimental conditions (C1–C5), participants were assigned to one of five combinations (Table 2). Each combination mapped one condition to each building, so that each participant experienced all five conditions, one per building. Using five distinct buildings – rather than repeating a single layout across conditions – ensured that each condition was experienced on a previously unseen building, preventing layout familiarity from carrying over between trials. Building order was randomized within participants to mitigate sequence effects.

Each layout was 3D-modeled in Rhino/Grasshopper and imported into Unity. Lighting, materials, and camera parameters were configured to approximate realistic viewing conditions. Viewpoint locations were initialized from the AGP-derived number of viewpoints and then validated and adjusted in Isovists.org app to match the five condition placements. The



(a) B1. **(b)** B2. **(c)** B3.

(d) B4. **(e)** B5.

Figure 7 Final set of five building layouts used in the experiment. Layouts were selected for structural variety and full compatibility with all five experimental conditions. Manual refinements enhanced room distinctiveness and eliminated redundant viewpoints.

resulting viewpoint configurations were imported into Unity, where an automated script applied the five experimental conditions (assigning viewpoints, inserting landmarks, ordering views, and exporting 360° panoramic images). These panoramas served as the core visual stimuli.

For each trial, one correct layout was presented to participants alongside three visually similar alternatives. The alternatives were created by editing the target layout so that they would match its overall complexity and distinctive geometric cues while differing in spatial configuration (in most cases this was achieved by rearranging room positions). Orientation was held constant across alternatives to avoid rotational confounds. An example set of alternatives is shown in Figure 6. Each participant received a base amount of £3 and could earn up to £4 through performance-based bonuses (£0.80 per layout × 5 layouts).

3.6 Pilot-informed protocol refinement

An iterative refinement process was conducted to align participant behaviour with the intended construct of *layout understanding* (integration across multiple vistas rather than single-view feature matching). Two pilot phases were run to validate the platform and identify unintended strategies. A first pilot phase with $n = 18$ volunteers validated the technical setup and gathered usability feedback across the five layout–condition combinations. A second pilot phase on Prolific (54 recruited; 42 retained after the checks in Appendix A) evaluated how interface access and incentive framing shaped exploration.

In both pilots, the layout-choice screen could be accessed after any number of viewed panoramas. A performance-based bonus scheme was used, in which a base payment of £3 was provided, and a bonus of up to £4 was offered, corresponding to five layout trials, with up to £0.80 available per layout if the correct layout was selected. The first two views were free and additional views were associated with small deductions (e.g., £0.10 per additional view). A single optional revisit was permitted after all views had been seen, with a one-time deduction of £0.10, and an incorrect choice resulted in forfeiture of the bonus for that layout. Under this structure, frequent early decisions were observed: layouts were often selected after

only one or two panoramas, and decision times on the layout-choice screen were relatively large (median decision time = 31.5 s, IQR: 20.0–46.8). Interaction logs and participants' feedback indicated that decisions were often made via *feature matching*–memorizing local details and then excluding alternatives–rather than by integrating the full set of vistas into a coherent layout representation. Because the use of this strategy by some participants confounded the results, the pilot data were not used for hypothesis testing.

The final protocol was then fixed for hypothesis testing: the tutorial trial was redesigned so that the view distinguishing the two alternatives was shown first rather than second, discouraging participants from relying on the last-viewed panorama for feature matching; access to the layout-choice screen was locked until all panoramas in a trial had been viewed at least once; per-view penalties were removed in favour of a flat £0.80 bonus for a correct answer; and a 40 s decision window was imposed to limit prolonged feature-matching on the alternatives. Evidence that these changes shifted behaviour toward the intended integrative strategy is reported in Section 4.1.

4 Results

4.1 Effect of protocol refinements on exploration behavior

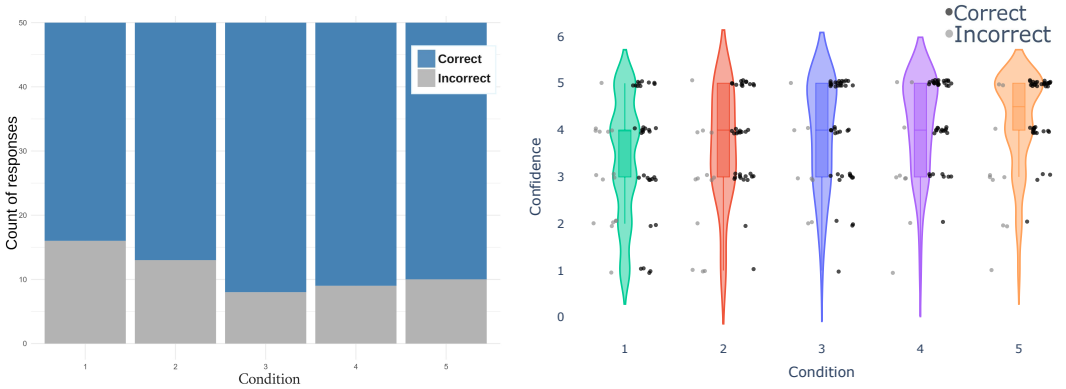
Before analysing task performance, we verified that the final protocol shifted exploration behaviour toward the intended integrative strategy and away from the single-view feature matching observed in the pilots. Median per-panorama viewing time fell from 23.9 s (IQR: 14.5–44.5) in the pilot to 13.7 s (IQR: 8.7–22.9) in the final protocol, and time spent on the final panorama fell from 18.9 s (IQR: 11.6–36.7) to 10.4 s (IQR: 7.2–17.3), consistent with reduced reliance on memorising isolated views, and on the last view in particular. Median decision time on the layout-choice screen also decreased, from 31.5 s (IQR: 20.0–46.8) to 23.5 s (IQR: 15.5–36.1) despite the 40 s cap, suggesting that participants selected the layout matching an already-formed representation rather than comparing features across alternatives. The optional revisit, available only after all views had been seen, was used 61 times across the 250 retained trials, compared with only 2 times in the pilots. Together, these patterns indicate improved compliance with the intended integrative strategy, and only data from the final protocol were retained for the reported analyses.

4.2 Descriptive overview of performance measures

We analysed three measures related to task performance: (i) correctness (objective accuracy), (ii) confidence ratings (subjective certainty), and (iii) an integrated confidence–correctness score.

Correctness. Correctness measures whether participants selected the correct layout from four options available and provides a direct indicator of task success. Figure 8a shows correct versus incorrect responses per condition (50 responses per condition; $N = 250$). Overall, 194 responses were correct. Accuracy was lowest in Condition 1 and higher in Conditions 3–5, suggesting that additional cues beyond baseline viewing may benefit recognition.

However, correctness alone does not distinguish between informed recognition and lucky guesses. Two participants may both answer correctly while differing substantially in certainty and depth of understanding, motivating the confidence ratings examined next.



(a) Correct vs. incorrect responses by condition. **(b)** Confidence ratings by condition.

Figure 8 Descriptive performance summaries across conditions (50 responses per condition).

Confidence. Confidence ratings provide a metacognitive estimate of participants' certainty in their layout judgments, potentially distinguishing informed decisions from guesses, thereby reflecting varying levels of layout understanding. Figure 8b shows the distribution of confidence ratings by condition. Consistent with the correctness pattern, Conditions 3–5 exhibit higher confidence distributions, whereas Condition 1 shows broader and lower confidence, indicating greater uncertainty.

To evaluate whether confidence meaningfully tracks performance in this task, calibration was assessed using a Confidence–Accuracy Characteristic (CAC) analysis. CAC plots, widely used to assess the diagnostic value of confidence, relate confidence levels to the proportion of correct responses at each level [23, 35, 48, 45]. As shown in Figure 9, accuracy increased with confidence overall. The small dip at confidence level 2 coincides with the smaller number of responses at that level. This confidence–accuracy alignment suggests that participants' confidence ratings tracked response accuracy overall and were therefore informative for distinguishing stronger from weaker layout understanding.

A limitation of raw confidence is that it does not reflect whether certainty is justified: high confidence after an incorrect choice corresponds to misplaced certainty and should not be interpreted as strong comprehension. To better capture meaningful understanding across conditions, a combined metric was therefore used.

Combined confidence–correctness score. To integrate subjective certainty with objective correctness, a combined score was computed:

$$\text{Combined Score} = \begin{cases} \text{Confidence rating (1–5)} & \text{if the answer is correct} \\ 0 & \text{if the answer is incorrect.} \end{cases}$$

This yields a six-level ordinal measure (0–5), where higher values reflect correct and confident recognition. The combined score aligns closely with the study goal of capturing *successful and justified* layout understanding, and was therefore used as the primary dependent variable in the main model. A correctness-only model is also reported as a robustness check.

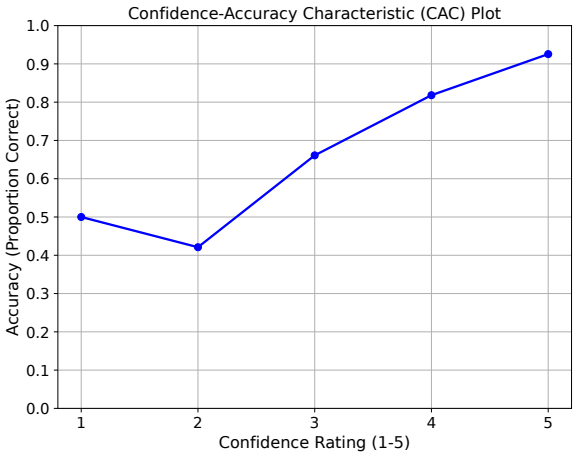


Figure 9 Confidence–accuracy characteristic (CAC) plot. Each point shows the proportion of correct responses at a given confidence level (1–5).

4.3 Modeling approach

Two complementary models were fitted in R [41] to quantify effects of viewing condition and spatial ability: (i) a mixed-effects logistic regression on correctness (robustness check), and (ii) a cumulative link mixed model on the combined confidence–correctness score suitable for ordinal data (primary analysis). Spatial ability was operationalized by the standardized PTT score (`PTT_z`).

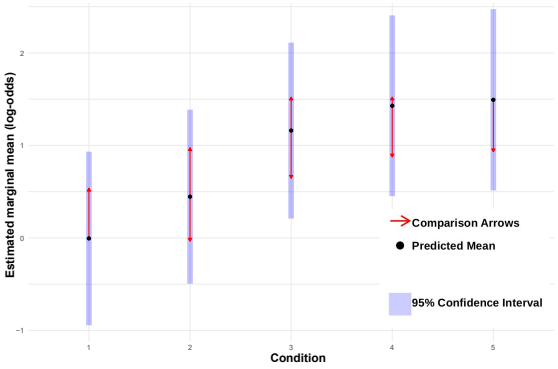
4.3.1 Robustness check: correctness model (binary DV)

Correctness was modeled as a binary outcome (0/1) using a mixed-effects logistic regression (`lme4` [2]) with fixed effects of `Condition` and `PTT_z`, and a random intercept for `Layout`:

$$\texttt{Correctness} \sim \texttt{Condition} + \texttt{PTT_z} + (1 \mid \texttt{Layout}). \tag{1}$$

A random intercept for Participant was initially included but was removed from the final model: once spatial ability (`PTT_z`) was added as a fixed effect, the participant-level variance dropped to zero (a singular fit), indicating that individual differences in correctness were fully captured by `PTT_z`. In this model, Spatial ability significantly predicted correctness (better spatial ability corresponding to higher accuracy; $p < .001$). In contrast, condition differences were not statistically significant after adjustment (all contrasts $p > .17$). Thus, although descriptive differences were visible, correctness alone did not yield strong inferential differences between conditions. Moreover, correctness does not capture participants' internal certainty: correct answers obtained through guessing are indistinguishable from correct answers based on confident understanding. This motivates the combined-score model as the primary analysis.

A post-hoc simulation-based power analysis (500 iterations) using the `simr` package [14] indicated that achieved power was 99.8% for spatial ability but only 42.8% for the condition effect, showing that the binary outcome was underpowered to detect condition differences.



(a) Estimated marginal means by condition.

Contrast	Estimate (log-odds)	SE	z-ratio	p-value (Dunnett)
H1: C1 as reference (testing if overlap is essential)				
C1 vs C2	0.452	0.384	1.177	0.5703
C1 vs C3	1.166	0.388	3.007	**0.0099**
C1 vs C4	1.434	0.403	3.562	**0.0014**
C1 vs C5	1.499	0.406	3.696	**0.0009**
H2a,b: C2 as reference (testing route sequencing and landmarks)				
C2 vs C3	0.715	0.384	1.860	0.1965
C2 vs C4	0.983	0.397	2.473	**0.0474**
C2 vs C5	1.047	0.400	2.615	**0.0322**
H2c: C5 as reference (testing additive benefit of both cues)				
C5 vs C3	-0.333	0.395	-0.841	0.7783
C5 vs C4	-0.065	0.403	-0.160	0.9954

(b) Planned condition contrasts (log-odds).

Figure 10 Model-based effects for the combined confidence–correctness score. (a) Estimated marginal means (points) with 95% CIs (purple bands); red arrows aid pairwise comparison (less overlap between arrows = stronger differences), and higher values indicate stronger layout understanding. (b) Condition contrasts (log-odds): 0 = no difference; positive values favour the second condition in each label, negative the first; bold p-values are significant.

4.3.2 Primary model: combined score (ordinal DV)

The primary analysis modeled the combined confidence–correctness score (0–5) using a cumulative link mixed model (logit link), fitted with the `ordinal` package [7], with fixed effects of `Condition` and `PTT_z` and random intercepts for `Participant` and `Layout`:

$$\texttt{CombinedScore} \sim \texttt{Condition} + \texttt{PTT_z} + (1 \mid \texttt{Participant}) + (1 \mid \texttt{Layout}). \tag{2}$$

Spatial ability was a significant predictor in this model ($p < .001$), indicating that participants with better spatial ability tended to achieve higher combined scores.

For the condition effect, estimated marginal means (Figure 10a) show an increasing trend from Condition 1 toward Conditions 3–5. Post hoc pairwise comparisons, computed with `emmeans` [27] using Dunnett adjustment,[2] (Figure 10b) showed that Conditions 3, 4, and 5 significantly outperformed Condition 1, while Condition 2 did not ($p = .57$). This pattern suggests that maximal viewpoint overlap alone (Condition 2) does not substantially enhance layout understanding. Both landmarks (C4) and their combination with route-based sequencing (C5) led to significantly higher scores compared to Condition 2 ($p = .047$ and $p = .032$, respectively), whereas route-based sequencing alone (C3 vs. C2) did not yield a significant improvement ($p = .20$). Condition 5 did not differ significantly from either Condition 3 or Condition 4, indicating that combining both cues produced no additive benefit beyond the presence of a single cue.

Revisit behavior. Revisits were most frequent in Condition 1 (18 revisits, with approximately 90% involving revisiting all available views), whereas other conditions ranged between 8–12 revisits with an average revisit extent of approximately 74%. This pattern is consistent with higher uncertainty or integration demand in the baseline condition.

[2] Dunnett adjustment accounts for multiple comparisons by testing each condition against a designated reference. In this experiment, C1, C2, and C5 were each used as reference levels to test H1 and H2a–c.

A post-hoc power analysis for the combined-score model yielded 98.4% power for the condition effect. Data were simulated (500 iterations) from the fitted model parameters – fixed effects, thresholds, and random-effect variances – and each iteration was evaluated with a likelihood-ratio test, confirming that this ordinal measure was well-powered to detect condition differences.

5 Discussion

This study examined how viewpoint-system design influences *layout understanding* from a minimal set of 360° panoramas. Motivated by the Network of Reference Frames theory [33], we varied three factors that can support spatial integration across discrete vista spaces: visual overlap (to enable correspondence between consecutive views), route-based ordering (to approximate continuity), and landmarks (to provide stable anchors). Using a combined confidence–correctness score as the primary outcome (with a correctness-only robustness check), we interpret the results in relation to the hypotheses and discuss theoretical, applied, and methodological implications.

Across analyses, spatial ability was a strong predictor of performance: participants with lower `PTT_z` (better perspective-taking ability) achieved higher correctness and higher combined scores, supporting H3. This is consistent with prior work showing that spatial integration efficiency varies across individuals and depends on the ability to form and use perspective shifts [18, 33, 34]. Participants with better perspective-taking ability may therefore have been better able to align viewpoints and integrate multiple vistas into a coherent layout representation.

Condition differences were not statistically significant in the correctness-only model, but were detectable once confidence was incorporated via the combined score. This is consistent with the post-hoc power analyses reported above: on the same sample, the binary correctness outcome afforded low power for condition effects (42.8%), whereas the combined score afforded high power (98.4%), indicating that the coarser binary measure limited sensitivity to condition differences. Taken together, the combined-score results suggest the following patterns:

- **Overlap alone is insufficient.** Condition 2, which maximized visual overlap without route-based sequencing or landmarks, did not significantly outperform the baseline Condition 1. Performance improved only when overlap was combined with additional cues, as in Conditions 3, 4, and 5. This suggests that visual overlap alone does not guarantee integrative spatial understanding: it appears to provide a foundation that the additional cues build on, rather than being effective on its own. H1 is therefore only partially supported.
- **Landmarks, not route-based sequencing, aid integration.** Conditions 4 and 5, which included landmarks, significantly outperformed Condition 2, whereas Condition 3, which added route-based sequencing without landmarks, did not. This suggests that, once overlap is available, landmarks rather than route-based sequencing are the cue that improves layout understanding, supporting H2b while not supporting H2a.
- **No additive benefit from combining cues.** Condition 5, which combined both landmarks and route-based sequencing, did not significantly outperform either Condition 3 or Condition 4. This indicates that, once overlap is available, the presence of either cue may be sufficient, and combining them does not yield additional gains. Thus, H2c is not supported.

Behavioral evidence was consistent with this interpretation. Revisit behavior was highest in the baseline condition (Condition 1), indicating higher uncertainty and/or higher integration demand in the least supportive configuration. Revisiting can be interpreted as an attempt to refine perspective shifts through repeated exposure, which has been shown to reduce integration costs [36, 33, 34].

These results offer several theoretical contributions. The Network of Reference Frames theory was used to inspire the experimental design and to explain the mechanism underlying participants' performance, rather than being tested as a theory. Within this scope, the results are consistent with the spatial-integration mechanism implied by the Network of Reference Frames theory: spatial knowledge is constructed by linking multiple vista spaces via perspective shifts, and both task design (viewpoint cues) and individual ability (spatial perspective-taking) influence how effectively this integration occurs [33, 34]. More specifically, this study extends this theory-informed line of work by operationalizing key integration-related ideas in a controlled panoramic-viewpoint design.

Visual overlap can enable correspondences between consecutive views, but it was not sufficient on its own to reliably improve layout understanding. This supports the idea that integration costs are not only a matter of whether overlap exists, but also whether reference-frame relations can be maintained consistently across transitions [34]. Landmarks likely reduced integration costs by providing stable, re-identifiable anchors within overlap regions, thereby strengthening perspective shifts and helping participants maintain a consistent orientation across views. This interpretation is consistent with prior findings that landmarks and allocentric cues support cross-view alignment and spatial memory precision [1, 15, 25].

The weak incremental effect of route-based ordering in the current layouts suggests boundary conditions. A route-like sequence may not yield strong benefits when views remain discrete and the environment is small (4–6 viewpoints), because ordering alone may not provide enough continuity to support sustained updating across transitions. This does not imply that movement-related continuity is unimportant in general; rather, it suggests that the route-based ordering used here may have been too weak to reproduce the benefits observed in VR locomotion research, where continuous movement or enriched transitions can improve spatial updating [31, 42, 3]. At the same time, our findings complement that literature by shifting the focus from *how transitions are rendered* to *how a minimal set of viewpoints is placed and sequenced*.

From an applied perspective, for panoramic viewpoint systems intended to support building understanding, the results suggest that maximizing overlap alone is unlikely to be sufficient. Adding structured cues – particularly landmarks that persist across overlapping vistas – is a practical way to improve users' comprehension. More broadly, this work provides a reproducible experimental pipeline for systematically testing different viewpoint-system configurations under controlled constraints and with fine-grained behavioral logging. This offers an initial step toward evidence-based design rules for panoramic tours and related spatial-learning interfaces.

Beyond the design factors, the refinement process showed how strongly task framing shapes exploration: the per-view penalty used in the pilots encouraged early decisions, whereas the flat reward with enforced full exploration promoted more integrative viewing (Section 4.1). This underscores the importance of carefully calibrated framing in online spatial-cognition studies, where participants may otherwise adopt confounding shortcut strategies.

5.1 Limitations

While the combined confidence–correctness score provided an informative primary outcome, it collapses all incorrect trials to zero regardless of the reported confidence. As a result, the score does not preserve potentially informative differences among errors (e.g., low-confidence mistakes versus high-confidence mistakes). Although the CAC analysis showed that confidence generally tracked correctness, high-confidence errors can still occur, reflecting overconfidence or "lucky" responding. Further refinement of scoring methods is therefore warranted. For example, calibration-aware measures such as the Brier Score [5] or Confidence-Based Marking (CBM) [13] explicitly reward alignment between confidence and correctness while penalizing miscalibrated responses. Incorporating such measures (either analytically or within the incentive structure) could encourage more cautious and better-calibrated judgments and provide a fuller picture of performance.

Layout understanding was measured through a multiple-choice layout recognition test. This format is objective and scalable, but it may not fully reflect participants' internal spatial representations. As observed in the pilot phases, some participants developed response strategies based on the characteristics of the choice alternatives themselves (e.g., feature matching and elimination) rather than the spatial information gathered from the viewpoints. Moreover, the same selected layout may reflect different reasoning strategies, depending on how similar the distractor layouts are to the correct layout. Although the final protocol was designed to reduce such shortcut strategies (e.g., unlocking the choice screen only after all views were explored and limiting decision time), performance in a multiple-choice format remains partly dependent on the distractor set. Complementary assessment formats – such as freehand sketching [44] or spatial reconstruction tasks – could provide deeper insight into participants' mental models, at the cost of additional scoring complexity of final results.

The tested environments were small single-floor layouts with relatively few viewpoints. This may have limited the observable benefit of route-based ordering: in larger or multi-floor environments with longer paths and more transitions, continuity cues may play a stronger role in supporting integration.

Relatedly, the five buildings were selected to be structurally distinct rather than matched on complexity, but no quantitative complexity measure was applied. Prior work shows that structural or geometric layout complexity can affect the spatial knowledge acquired [38], so differences in complexity across the five buildings may have contributed to the between-layout variability in performance. This variability was absorbed by the layout random intercept in all models, so it did not bias the condition or spatial-ability estimates; nonetheless, an explicit complexity measure – for example, based on isovist or space-syntax properties – would allow this variability to be characterised rather than only controlled for, and is a worthwhile direction for future stimulus design.

Finally, prior experience with virtual environments or video games was not collected, yet such experience may influence comfort with panoramic navigation and exploration efficiency, and could thus contribute to individual differences independently of spatial ability; future studies could include it as a covariate. In addition, the sample was not balanced by gender (34 male, 16 female), which may limit the generalisability of the findings.

5.2 Future work

Several directions can extend this work and address its limitations. One open question concerns *orientation alignment*. Because initial panorama headings were randomized, participants may have incurred additional alignment costs at each view. Misaligned orientations

are known to increase integration costs; for example, Meilinger [34] reported substantially higher errors when integrating information across 90°-rotated viewpoints compared to aligned ones. Future studies can compare (i) consistent global orientation (e.g., "north-up" initialization), (ii) alignment toward the next overlap region, and (iii) adaptive initialization based on the preceding viewpoint, to test whether reducing alignment disparity improves integration and confidence.

Another promising direction is to leverage the hierarchical structure of spatial memory. Hierarchical theories of spatial cognition propose that not all vista spaces are equally represented in memory [33, 19, 47]. Smaller or peripheral spaces may be encoded at lower hierarchical levels, while more central or integrative spaces may serve as higher-level anchors. A key rationale behind this structuring lies in clustering effects – where spatial representations are organized not as isolated views but as grouped regions or spatial units. Rather than presenting all views as independent and equally important, future studies could present spatially clustered views as cohesive groups, emphasizing the transitions between these higher-order clusters instead of transitions between individual views. For example, a central hall and its adjacent rooms could be presented as one group, followed by other overlapping or connected groups. One possible basis for defining such groups is the concept of *m-spaces* [40], i.e., to the minimal set of convex spaces required to cover an environment. A hierarchical presentation could then begin with the most connected m-spaces, followed by successively linked clusters, reflecting how spatial knowledge may be naturally organized in memory.

6 Conclusion

This exploratory study examined how viewpoint-system design supports building-layout understanding from a minimal set of 360° panoramas. Building on the Network of Reference Frames theory and spatial-integration concepts, three design factors were systematically manipulated: visual overlap between consecutive views, route-based sequencing, and landmark availability. A reproducible, web-based experimental framework was developed to implement these manipulations under controlled conditions and to log exploration behavior and responses in detail.

The results indicate that individual spatial ability – specifically *perspective-taking* – was a strong predictor of layout understanding ($p < .001$). In addition, viewpoint design was found to modulate performance: high visual overlap alone did not reliably improve performance, meaning that a hypothetical guard from the Art Gallery Problem solution would be unlikely to understand the gallery layout even with the help of overlapping views. Yet, overlap supported layout understanding as soon as it was paired with additional cues. Among the tested cues, landmarks produced the clearest improvement in layout understanding ($p < .05$). Route-based sequencing showed weaker effects in the present setting, and combining landmarks with route-based sequencing did not yield additional benefits beyond either cue alone.

Overall, these findings suggest a simple design implication for panoramic tours: overlap is helpful, but not sufficient on its own; stable landmarks visible across overlapping views can make layout understanding easier. At the same time, a central contribution of this work is the reproducible experimental pipeline. Because the framework is modular, new design variations (e.g., stronger continuity cues, different orientation rules, or alternative evaluation tasks) can be implemented quickly, tested efficiently, and compared under the same controlled setup. This provides a practical foundation for future studies to test additional design factors and to develop more general, evidence-based rules for generating effective viewpoint systems.

References

1 Christina Adamou, Marios N. Avraamides, and Jonathan W. Kelly. Integration of visuospatial information encoded from different viewpoints. *Psychonomic Bulletin & Review*, 21(3):659–665, 2014. doi:10.3758/s13423-013-0538-5.

2 Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi:10.18637/jss.v067.i01.

3 Jiwan Bhandari, Paul MacNeilage, and Eelke Folmer. Teleportation without spatial disorientation using optical flow cues. In *Proceedings of the 44th Graphics Interface Conference*, GI '18, pages 162–167, Waterloo, CAN, 2018. Canadian Human-Computer Communications Society. doi:10.20380/GI2018.22.

4 Sarah E Bonner and Geoffrey B Sprinkle. The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society*, 27(4):303–345, 2002. doi:10.1016/S0361-3682(01)00052-6.

5 Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

6 Lucia A. Cherep, Alex F. Lim, Jonathan W. Kelly, Devi Acharya, Alfredo Velasco, Emanuel Bustamante, Alec G. Ostrander, and Stephen B. Gilbert. Spatial cognitive implications of teleporting through virtual environments. *Journal of Experimental Psychology: Applied*, 26(3):480–492, 2020. doi:10.1037/xap0000263.

7 Rune H. B. Christensen. *ordinal—Regression Models for Ordinal Data*, 2023. R package version 2023.12-4.1. URL: https://CRAN.R-project.org/package=ordinal.

8 V. Chvátal. A combinatorial theorem in plane geometry. *Journal of Combinatorial Theory, Series B*, 18:39–41, 1975. doi:10.1016/0095-8956(75)90061-1.

9 Ruth Conroy. *Spatial navigation in immersive virtual environments*. Phd thesis, University College London, 2001. URL: https://discovery.ucl.ac.uk/id/eprint/1111.

10 Joshua R. de Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12, 2015. doi:10.3758/s13428-014-0458-y.

11 Joshua R. de Leeuw. Datapipe: Born-open data collection for online experiments. *Behavior Research Methods*, 56(3):2499–2506, 2024. doi:10.3758/s13428-023-02161-x.

12 Alinda Friedman, Bernd Kohler, Peri Gunalp, Alexander P. Boone, and Mary Hegarty. A computerized spatial orientation test. *Behavior Research Methods*, 52(2):799–812, 2020. doi:10.3758/s13428-019-01277-3.

13 A. R. Gardner-Medwin. Confidence assessment in the teaching of basic science. *Research in Learning Technology*, 3(1), 1995. doi:10.3402/rlt.v3i1.9597.

14 Peter Green and Catriona J. MacLeod. Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016. doi:10.1111/2041-210X.12504.

15 Nathan Greenauer, Catherine Mello, Jonathan W. Kelly, and Marios N. Avraamides. Integrating spatial information across experiences. *Psychological Research*, 77(5):540–554, 2013. doi:10.1007/s00426-012-0452-x.

16 Nathan Greenauer and David Waller. Micro- and macroreference frames: Specifying the relations between spatial categories in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36:938–957, 2010. doi:10.1037/a0019647.

17 Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004. doi:10.1016/j.intell.2003.12.001.

18 Mary Hegarty and David A. Waller. Individual differences in spatial abilities. In Priti Shah and Akira Miyake, editors, *The Cambridge Handbook of Visuospatial Thinking*, pages 121–169. Cambridge University Press, Cambridge, 2005. doi:10.1017/CBO9780511610448.005.

19 Stephen C. Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory and Cognition*, 13(3):208–217, 1985. doi:10.3758/BF03197683.

20 Corinne A. Holmes, Steven A. Marchette, and Nora S. Newcombe. Multiple views of space: Continuous visual flow enhances small-scale spatial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(6):851–861, 2017. doi:10.1037/xlm0000346.

21 Ronald Hübner and Jan Schlösser. Monetary reward increases attentional effort in the flanker task. *Psychonomic Bulletin & Review*, 17(6):821–826, 2010. doi:10.3758/PBR.17.6.821.

22 Toru Ishikawa and Daniel R. Montello. Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, 52(2):93–129, 2006. doi:10.1016/j.cogpsych.2005.08.003.

23 Peter Juslin, Nils Olsson, and Anders Winman. Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5):1304–1316, 1996. doi:10.1037/0278-7393.22.5.1304.

24 Amy A Kalia, Gordon E Legge, and Nicholas A Giudice. Learning building layouts with non-geometric visual information: The effects of visual impairment and age. *Perception*, 37(11):1677–1699, 2008. doi:10.1068/p5915.

25 Jonathan W. Kelly and Timothy P. McNamara. Reference frames during the acquisition and development of spatial memories. *Cognition*, 116(3):409–420, 2010. doi:10.1016/j.cognition.2010.06.002.

26 Maria Kozhevnikov and Mary Hegarty. A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & cognition*, 29(5):745–756, 2001. doi:10.3758/BF03200477.

27 Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2025. R package version 1.11.1. doi:10.32614/CRAN.package.emmeans.

28 Changyang Li, Haikun Huang, Jyh-Ming Lien, and Lap-Fai Yu. Synthesizing scene-aware virtual reality teleport graphs. *ACM Trans. Graph.*, 40(6), December 2021. doi:10.1145/3478513.3480478.

29 Hengshan Li, Panagiotis Mavros, Jakub Krukar, and Christoph Hölscher. The effect of navigation method and visual display on distance perception in a large-scale virtual building. *Cognitive Processing*, 22:239–259, 2021. doi:10.1007/s10339-020-01011-4.

30 Tingguang Li, Danny Ho, Chenming Li, Delong Zhu, Chaoqun Wang, and Max Q.-H. Meng. Houseexpo: A large-scale 2d indoor layout dataset for learning-based algorithms on mobile robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5839–5846, 2020. doi:10.1109/IROS45743.2020.9341284.

31 Alex F. Lim, Jonathan W. Kelly, Nathan C. Sepich, Lucia A. Cherep, Grace C. Freed, and Stephen B. Gilbert. Rotational self-motion cues improve spatial learning when teleporting in virtual environments. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction*, SUI '20, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3385959.3418443.

32 Jack M. Loomis, Roberta L. Klatzky, Reginald G. Golledge, and John W. Philbeck. Human navigation by path integration. In Reginald G. Golledge, editor, *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*, pages 125–151. Johns Hopkins University Press, Baltimore, MD, 1999.

33 Tobias Meilinger. The network of reference frames theory: A synthesis of graphs and cognitive maps. In Christian Freksa, Nora S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, pages 344–360. Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-87601-4_25.

34 Tobias Meilinger, Alain Berthoz, and Jan M. Wiener. The integration of spatial information across different viewpoints. *Memory & Cognition*, 39(6):1042–1054, 2011. doi:10.3758/s13421-011-0088-x.

35 Laura Mickes. Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, 4(2):93–102, 2015. doi:10.1016/j.jarmac.2015.01.003.

36 Daniel R. Montello. Spatial orientation and the angularity of urban routes: A field study. *Environment and Behavior*, 23(1):47–69, 1991. doi:10.1177/0013916591231003.

37 Daniel R. Montello. Scale and multiple psychologies of space. In Andrew U. Frank and Irene Campari, editors, *Spatial Information Theory A Theoretical Basis for GIS*, pages 312–321. Springer, 1993. doi:10.1007/3-540-57207-4_21.

38 Michael J. O'Neill. Evaluation of a conceptual model of architectural legibility. *Environment and Behavior*, 23(3):259–284, 1991. doi:10.1177/0013916591233001.

39 Joseph O'Rourke. *Art Gallery Theorems and Algorithms*. Oxford University Press, USA, 1987.

40 J Peponis, J Wineman, M Rashid, S Hong Kim, and S Bafna. On the description of shape and spatial configuration inside buildings: Convex partitions and their local properties. *Environment and Planning B: Planning and Design*, 24(5):761–781, 1997. doi:10.1068/B240761.

41 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL: https://www.R-project.org/.

42 Kasra Rahimi, Colin Banigan, and Eric D. Ragan. Scene transitions and teleportation in virtual reality and the implications for spatial awareness and sickness. *IEEE Transactions on Visualization and Computer Graphics*, 26(6):2273–2287, 2020. doi:10.1109/TVCG.2018.2884468.

43 Yugandhar Shinde, Kyeongsuk Lee, Beyza Kiper, Makayla Simpson, and Sogand Hasanzadeh. A systematic literature review on 360° panoramic applications in architecture, engineering, and construction (aec) industry. *Journal of Information Technology in Construction*, 28:405–437, 2023. doi:10.36680/j.itcon.2023.021.

44 M. Simonet, C. Vater, C. Abati, S. Zhong, P. Mavros, A. Schwering, M. Raubal, C. Hölscher, and J. Krukar. Probing mental representations of space through sketch mapping: a scoping review. *Cognitive Research: Principles and Implications*, 10(1):59, 2025. doi:10.1186/s41235-025-00667-w.

45 Alexander M Smith and Amy-May Leach. Confidence can be used to discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science*, 14(6):1062–1071, 2019. doi:10.1177/1745691619863431.

46 Dajana Snopková, Hana Švedová, Petr Kubicek, and Zdeněk Stachoň. Navigation in indoor environments: Does the type of visual learning stimulus matter? *ISPRS International Journal of Geo-Information*, 8, 2019. doi:10.3390/ijgi8060251.

47 Albert Stevens and Patty Coupe. Distortions in judged spatial relations. *Cognitive Psychology*, 10(4):422–437, 1978. doi:10.1016/0010-0285(78)90006-3.

48 Eylul Tekin, K. Andrew DeSoto, John H. Wixted, and Henry L. Roediger III. Applying confidence accuracy characteristic plots to old/new recognition memory experiments. *Memory*, 29(4):427–443, 2021. doi:10.1080/09658211.2021.1901937.

49 Perry W. Thorndyke and Barbara Hayes-Roth. Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, 14(4):560–589, 1982. doi:10.1016/0010-0285(82)90019-6.

50 Davi C. Tozoni, Pedro J. de Rezende, and Cid C. de Souza. Algorithm 966: A practical iterative algorithm for the art gallery problem using integer linear programming. *ACM Trans. Math. Softw.*, 43(2), 2016. doi:10.1145/2890491.

A Participant exclusion criteria

Exclusion criteria were predefined and applied conservatively, so that a participant was excluded only when multiple independent indicators of inattentiveness or non-compliance co-occurred. General checks across all phases removed participants who used a mobile device, submitted incomplete data (fewer than five output files), or restarted the experiment and submitted duplicate files; none were excluded on these grounds.

In the second pilot phase, where the layout-choice screen could be opened after any number of viewed panoramas, exclusion was based on the following indicators. A *nonsensical-confidence* failure occurred when a participant decided after a single view in the tutorial while reporting very high confidence: because the tutorial was unsolvable from the first view alone, and participants who attempted to decide after the first view were explicitly warned of this, yet could still proceed. Reporting very high confidence in this situation flagged the participant as providing nonsensical responses, and led to exclusion on its own.

Three further indicators were defined: (i) deciding after a single tutorial panorama without high confidence, (ii) viewing fewer than three panoramas across all five trials – both reflecting consistent guessing or feature matching rather than layout integration – and (iii) skipping two or more of the three instruction screens (viewing time below a minimum threshold), reflecting inattention. A participant who met at least two of these three indicators was also excluded. In total, 12 of 54 participants were excluded.

In the final protocol, full exploration of all panoramas was enforced by the interface, so only two checks remained: below-threshold instruction-screen viewing times (as above) and a total task duration under five minutes. Participants were excluded only if they met both; 4 of 54 were excluded, yielding the final analytic sample of 50.

Much of Geospatial Web Search Is Beyond Traditional GIS

Ilya Ilyankou[1]
SpaceTimeLab, Department of Civil, Environmental, and Geomatic Engineering, University College London, UK

Stefano Cavazzi
Ordnance Survey, Southampton, UK

James Haworth
SpaceTimeLab, Department of Civil, Environmental, and Geomatic Engineering, University College London, UK

Abstract

Web search queries concern place far more often than existing labelling schemes suggest, yet the landscape of geospatial web search queries – what people ask of place, and how often – remains poorly characterised at scale. We apply dense sentence embeddings, a lightweight SetFit classifier, and density-based clustering to the full MS MARCO corpus of 1.01 million real Bing queries without prior filtering for toponyms or spatial keywords, identifying 181,827 geospatial queries (18.0%), nearly threefold the 6.17% labelled as *Location* in the original annotations. The resulting taxonomy of 88 query categories reveals that geospatial web search is dominated by transactional and practical lookups: costs and prices alone account for 15.3% of geospatial queries, nearly twice the size of the entire physical geography theme. Much of this activity – costs, opening hours, contact details, weather, travel recommendations – falls outside the scope of what traditional GIS and knowledge graphs are built to serve. The categories vary substantially in the kind of answer they admit, from deterministic lookups answerable from spatial databases or knowledge graphs to evaluative or temporally volatile queries that require generative or real-time systems. We discuss implications for hybrid retrieval architectures and for benchmarks of geographic reasoning in large language models. We openly release the labelled dataset, classifier, and taxonomy.

2012 ACM Subject Classification Information systems → Web searching and information discovery; Information systems → Geographic information systems; Information systems → Web log analysis; Computing methodologies → Natural language processing; Computing methodologies → Cluster analysis; Human-centered computing → Human computer interaction (HCI)

Keywords and phrases Web search queries, geographic information retrieval, query classification, geospatial query taxonomy, MS MARCO, sentence embeddings, density-based clustering, GeoAI, large language models, place theory

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.10

Related Version *Full Version*: `https://arxiv.org/abs/2605.11336`

Supplementary Material *Software (Code)*: `https://github.com/ilyankou/ms-marco-geospatial` archived at `swh:1:dir:67b243446916b179b6f0a39c61afe70a2fa5f4f0`
Model (Geospatial query classifier): `https://huggingface.co/ilyankou/is-geospatial-query`

Funding *Ilya Ilyankou*: This work was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1].

[1] Corresponding author

© Ilya Ilyankou, Stefano Cavazzi, and James Haworth;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 10; pp. 10:1–10:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Place is central to how people make sense of the world, yet existing information systems represent it in an impoverished way, often reducing it to named points or coordinate-bound objects rather than the vague, relational, and socially constructed phenomenon it actually is [34, 10, 18]. Large-scale web search query logs offer an empirical window into what people want from place: whether a town falls in a particular county, what the weather or cost of living is somewhere, where the nearest airport is, or what language is spoken in a country. The MS MARCO corpus of 1.01 million real Bing queries [5] remains the largest publicly available record of such behaviour, despite being collected prior to 2018 and skewed toward Anglophone North American users. It captures web search rather than conversational, voice, or professional GIS use, but it is the most suitable corpus currently available for characterising geospatial web search at scale.

Prior work has identified geospatial queries primarily by syntactic form or the presence of toponyms [17, 18, 25], producing prevalence estimates that are typically too low and taxonomies too narrow to capture the full range of what people want from place. We argue that the right starting point is the query corpus itself, without prior filtering. By applying Transformer sentence embeddings [36] and density-based clustering to the full MS MARCO dataset, we build a data-driven taxonomy of geospatial web search queries that quantifies not just how many queries are geospatial, but what categories they fall into and how those categories are distributed. Much of this activity, such as costs, opening hours, travel recommendations, or weather, falls outside the scope of what traditional GIS is built to serve, with direct implications for hybrid retrieval and generative system design.

We address three research questions:

1. What proportion of web search queries are geospatial, under a definition broader than "contains toponyms"?
2. What are the dominant themes of geospatial web search, and how are they distributed?
3. How do the resulting categories vary in the kind of answer they admit, and what does this imply for the design of systems that serve such queries?

Our work makes three practical contributions. First, we release a gold human-annotated dataset of 1,200 web search queries, labelled as geospatial or non-geospatial, constructed to cover the full embedding space of MS MARCO rather than its high-density regions. Second, we train and release a lightweight binary classifier that identifies geospatial queries with $F_1 = 0.930$ (95% bootstrap CI $0.909 - 0.947$) and can be applied to web search query corpora beyond MS MARCO. Third, we derive and release a taxonomy of 88 geospatial query categories, illustrated in Figure 3, that quantifies the relative prevalence of distinct query types and provides a concrete empirical characterisation of geospatial web search at scale.

We make the code available on GitHub[2] under the MIT license for reproducibility.

2 Related work

2.1 Taxonomies of geographic questions

Hamzei et al. mined MS MARCO [5] to characterise place questions by place type and scale [17], and later by a broader semantic schema covering place names, types, activities, spatial relationships, and qualities [16]. Xu et al. showed that professional geographic questions drawn from GIScience literature are syntactically richer and semantically distinct

[2] `https://github.com/ilyankou/ms-marco-geospatial`

from web queries, centring on analytical entities such as distribution, pattern, and density [42]. Another top-down approach by Kuhn et al. derived question templates from a spatial ontology and a taxonomy of place facets, arguing that corpus-driven methods inevitably under-represent emotive and physical facets due to collection bias [25].

On the system side, Punjani et al. addressed answering geographic natural language questions over structured linked data, translating questions into GeoSPARQL queries via handcrafted templates. Their GeoQuestions201 benchmark defines seven question categories of increasing structural complexity, covering topological, cardinal direction, and distance relations [33]. Kefalidis et al. extended this line of work with GeoQuestions1089, a larger benchmark, and showed that even improved template-based systems leave substantial headroom on geographic question answering [24].

2.2 Geospatial content in search queries and on the web

Several studies have attempted to quantify how much of online information and search activity is geographic in nature, using methods that range from keyword matching to network analysis to LLM-based classification.

Sanderson and Kohler found that roughly 18.6% of 1 million Excite[3] search queries studied contained a geographic term, with spatial relationship terms appearing rarely, suggesting users at the time did not expect search engines to handle relational spatial queries [38]. Jones et al. estimated place-name prevalence of Yahoo queries at 12.7%; they found that the vast majority target cities, and demonstrated that acceptable query-to-target distance is strongly topic-dependent, with restaurant queries clustering within tens of kilometres while accommodation queries tolerate far greater distance [23]. Gan et al. analysed the AOL dataset of 36 million records from 2006[4]; the authors built a geo/non-geo classifier and identified 13.39% of geographic-intent queries, and suggested a taxonomy of 23 broad and high-level categories that "combine aspects of topicality and desired type of interaction", such as "Tourism/Travel", "Medical", "Entertainment", or "Navigational" [12]. Henrich and Lüdecke also classified AOL user intent, and found habitation, accommodation, spare time, and information as dominant concepts; 65% of queries implied physical travel to the target; and the majority were selective rather than covering, seeking a specific point rather than documents spanning an area [19].

Beyond queries, Hahmann and Burghardt tested the much-cited claim that 80% of all information is geospatially referenced, finding via network and cognitive analyses of the German Wikipedia that the defensible figure is closer to 56–59% [15]. A more recent analysis of Common Crawl found that 18.7% of web documents contain explicit geospatial information such as coordinates or addresses [22], remarkably close to Sanderson and Kohler's query-side estimate two decades earlier.

2.3 Place theory and information needs

Edwardes and Purves showed that spatial vocabularies are strongly locale-sensitive; British contributors preferred *hill* to *mountain*, and human settlements dominated over physical geography, implying that any fixed keyword list will misclassify a substantial portion of implicitly spatial queries [11]. Purves et al. argued more fundamentally that existing information systems represent place in an impoverished way, reducing it to named points or

[3] Major web portal and search engine of the late 1990s
[4] The dataset was withdrawn shortly after release following privacy concerns

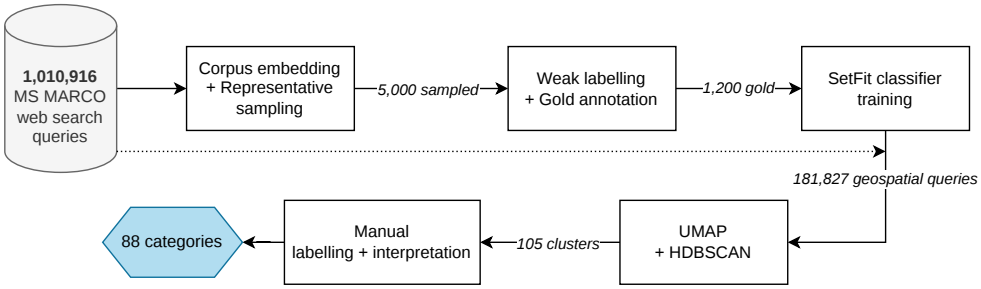


Figure 1 Overview of the methodology. Starting from the full MS MARCO corpus of 1.01M queries, we sample 5,000 representative queries for gold annotation, train a SetFit classifier to identify 181,827 geospatial queries, and apply UMAP dimensionality reduction and HDBSCAN clustering on geospatial query embeddings to derive a taxonomy of 88 geospatial query categories, grouped into 9 themes.

coordinate-bound objects, and that place is instead vague, relational, and socially constructed [34]. Therefore, it would follow that a query about living costs is just as place-anchored as any other query containing a toponym. Shanon [39] demonstrated that *where-question* responses are governed not only by containment hierarchies and physical distance but by the cognitive salience of referents and shared epistemic context, dimensions that template-driven geographic information retrieval systems do not fully address to this day.

2.4 Limitations of prior work

Previous query log studies and taxonomy building attempts share a critical limitation: they characterise geospatial queries by their syntactic form or semantic encoding, not by what people are actually trying to find out. Even sophisticated approaches to detecting geospatial language in text, such as spatial role labelling of geospatial prepositions [35], rely on surface-form signals and thus miss implicitly geospatial queries. Most rely on pre-filtered subsets: explicitly location-labelled records, toponym-containing queries, or purpose-built corpora that by design exclude such queries, producing prevalence estimates that are systematically too low when filtering on toponyms or place names (12.7% [23], 13.39% [12]), or that conflate surface form with intent when matching geographic terms more broadly. What remains unknown is the topical landscape of geospatial web search: not whether a query contains a place name or a spatial predicate, or how people phrase their queries, but whether people are searching for nearby restaurants, flight times between cities, the best places to move to, local weather, or regional costs of living. We address this gap by characterising geospatial queries by topical intent rather than syntactic form, applied without pre-filtering to a corpus of 1.01 million web search queries.

3 Methodology

Figure 1 summarises the full methodology, from corpus assembly through to taxonomy derivation.

3.1 Defining *geospatial*

No consensus definition of a spatial, geospatial, or geographic question or query exists. In the field of geographical question answering, Mai et al. broadly define *geographic* questions as those "involving geographic entities, geographic concepts, or spatial relations as parts of

the natural language questions" [27], and Kefalidis et al. define *geospatial* questions as those "requiring qualitative or quantitative geographic knowledge to be answered" [24]. We thus define a *geospatial query* as follows:

> A query is geospatial if it requires qualitative or quantitative geographic knowledge of Earth-bound features to be answered. This is usually the case if the query involves a geographic entity, a geographic concept, or a spatial relation.

In our definition, we exclude questions that are anatomical, microscopic, or astronomical in scale, and fictional or abstract "where" questions.

3.2 Identifying geospatial queries

We combined all available MS MARCO search queries[5] from the train, validation, and test subsets to produce a single corpus of 1,010,916 unique web search queries, each represented as a short[6], typically lowercased natural-language string, ranging from clearly not geospatial (e.g., "what is the human main muscles", *query ID 825350*), to clearly geospatial (e.g., "what county is badger mn in?", *602750*), to many contestable, such as "most western point in portugal" (*459663*), "how many square feet in an acre/" (*296354*), and "what all places do i need to change my address when i move" (*553148*).

To support sampling and downstream clustering, we pre-compute 384-dimensional embeddings for all 1.01M search queries using a compact, general-purpose text embedding model *BAAI/bge-small-en-v1.5*[7] via Sentence Transformers [36], which have been shown to encode some quasi-geospatial structure [21]. The chosen model offers strong semantic similarity performance for short English texts and is computationally feasible on a consumer laptop; the BGE family generally performs competitively on the Massive Text Embedding Benchmark (MTEB)[8] [32].

3.2.1 Constructing a gold dataset

We sample 5,000 queries from the entire MS MARCO corpus for labelling; exploratory manual verification suggested this would yield at least 500 positive (i.e., geospatial) samples. We use *k-means++*[9] [3] to generate 5,000 centroids in the embedding space and snap each to its nearest query. Compared to random sampling, we expect this to produce broader coverage of rare query types rather than over-representing dense clusters.

For each sampled query, we first obtain a *weak* label using a locally hosted Llama-3.1 [13] via Ollama[10] using the few-shot prompt shown in the Appendix A. The model is computationally efficient to run locally, and its instruction-following capability is sufficient for binary classification of short natural-language strings. To reduce the impact of stochastic variation in LLM outputs, we run the model 5 times per query at a low sampling temperature of 0.3 and assign the majority vote as the weak label. The lead author then manually verifies and corrects the weak labels to produce the gold dataset of $1,200$ queries, which we release

[5] Available to download at `https://msmarco.z22.web.core.windows.net/msmarcoranking/queries.tar.gz`
[6] Mean query length is 35 characters, and 99.9% of queries are under 131 characters
[7] Available on HuggingFace: `https://huggingface.co/BAAI/bge-small-en-v1.5`
[8] The leaderboard is available at `https://huggingface.co/spaces/mteb/leaderboard`
[9] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.kmeans_plusplus.html`
[10] `https://ollama.com/library/llama3.1`

on GitHub[11]. To assess the reliability of the lead author's gold labels, the remaining two authors independently of each other annotate a 200-query sample drawn from the gold set; we report pairwise Cohen's κ [9] for inter-annotator agreement across three annotator pairs.

3.2.2 Training a SetFit binary classifier

We use the gold dataset to contrastively train a lightweight geospatial/non-geospatial binary classifier using SetFit [41], a few-shot learning framework that fine-tunes a sentence embedding model via contrastive learning before fitting a classification head for strong performance with limited labelled data. We use *BAAI/bge-small-en-v1.5* as the base sentence-transformer. SetFit is designed for few-shot learning and requires few labelled examples to perform well [41]; we therefore split the dataset with stratification into train (200 samples), validation (200), and hold-out test (800) partitions for a robust initial evaluation. We train for five epochs with a batch size of 64 and a learning rate of 2×10^{-5}, keeping the SetFit default of 20 contrastive pair-sampling iterations per epoch. We evaluate every epoch and apply early stopping after 2 rounds with no improvements in embedding loss. The model is evaluated on the hold-out test set; we report 95% confidence intervals computed via 1,000 bootstrap resamples of the test set. For production inference (i.e., to label the 1.01M queries as geospatial or not), we retrain the same configuration on the full gold dataset for 3 epochs, corresponding to the best validation checkpoint observed during initial evaluation. We release the trained classifier on HuggingFace[12].

3.3 Building taxonomy

Once all geospatial queries are identified, we use clustering on their embeddings and manual interpretation to categorise geospatial web search and identify key themes.

3.3.1 Dimensionality reduction and density-based clustering

We take the subset of queries predicted as geospatial and retrieve their corresponding 384-d embeddings. Because density-based clustering in high dimensions is difficult (the problem often referred to as the *curse of dimensionality* [4]), we first reduce dimensionality using the Uniform Manifold Approximation and Projection algorithm, or UMAP[13] [29], which projects high-dimensional embeddings into a lower-dimensional space while preserving local neighbourhood structure, and then cluster the reduced representations using Hierarchical Density-Based Spatial Clustering of Applications with Noise, or HDBSCAN[14] [8]. This combination is widely used for topic modelling over transformer embeddings [2, 40, 14, 1].

3.3.2 Grid search over UMAP and HDBSCAN parameters

Density-based clustering relies on density thresholds whose appropriate values depend on the underlying data distribution; selecting a suitable density level is inherently difficult and data-dependent [7]. We therefore perform a grid search over an interpretable parameter space for both UMAP and HDBSCAN to identify a set of parameters that produces a reasonable number of high-quality clusters. For UMAP, we vary (i) output dimensionality $\in \{5, 10, 15\}$

[11] `https://github.com/ilyankou/ms-marco-geospatial`/tree/main/gold-dataset
[12] `https://huggingface.co/ilyankou/is-geospatial-query`
[13] `https://umap-learn.readthedocs.io/en/latest/`
[14] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html`

and (ii) neighbourhood size $\in \{10, 25, 50\}$. For HDBSCAN, we vary (iii) minimum cluster size $\in \{25, 50, 100, 200\}$ and (iv) minimum number of samples, set to $\{0.2, 0.5, 1.0\}\times$ *minimum cluster size*, using the default Excess of Mass (*"eom"*) cluster selection method. This gives us $3 \times 3 \times 4 \times 3 = 108$ configurations, each evaluated with a fixed random seed (42).

For each configuration, we compute four quality metrics: (a) the Density-Based Clustering Validation (DBCV) score [31], a density-aware measure of within-cluster cohesion and between-cluster separation suited to the arbitrary-shape clusters produced by methods such as HDBSCAN (range -1 to 1, higher is better), (b) the noise fraction, i.e. the proportion of queries assigned to cluster "-1"), (c) the number of non-noise clusters, and (d) the median non-noise cluster size. We use DBCV rather than silhouette scores, as the latter assumes convex clusters and has no principled treatment of noise points, both of which make it unsuitable for evaluating HDBSCAN output.

3.3.3 Consistency checks

UMAP is stochastic, and its randomness can affect downstream clustering. To assess how stable our clustering is, we pick the top-10 configurations by DBCV that produced at least 10 clusters (to exclude degenerate solutions with too few clusters to be informative), and re-run each across six random seeds $\in \{0, 1, 2, 3, 4, 42\}$. For each configuration, we report the mean and standard deviation of DBCV, noise fraction, and cluster count across seeds, and select the configuration with the strongest combination of high mean DBCV and low cross-seed variance. For final taxonomy extraction, we run the full pipeline once on all 181,827 queries using the selected configuration.

3.3.4 Cluster interpretation

To support manual naming of resulting clusters, for each cluster we (i) identify a representative query, defined as the query whose embedding is closest to the cluster centroid, (ii) extract salient unigram and bigram terms using Term Frequency-Inverse Document Frequency (TF-IDF) over concatenated cluster texts, with English stop-words removed, and (iii) randomly sample 10 queries per cluster to provide additional context for annotators. These summaries are exported as Markdown and as a spreadsheet for manual annotation by the authors.

Following manual annotation, we merge clusters that we judge to represent the same query intent (most notably, multiple US state-specific "what county is [place] in" clusters fragmented by named entities rather than intent), yielding a final taxonomy of 88 categories.

We group the categories into nine broad themes – Statistical, Temporal, "POIs and Commercial", Administrative, Physical, Cultural, Historical, Biographical, and Events – for organisational convenience rather than as a theoretically grounded hierarchy. We experimented with alternative groupings but found that meaningful structure resides in the leaf categories themselves; any higher-level grouping risks implying a cleaner taxonomy than the data supports.

The resulting taxonomy is exported as a parent–child JSON graph for visualisation.

4 Results

4.1 Gold dataset

Of the 5,000 sampled queries, 628 received at least one "geospatial" vote from the LLM, of which the vast majority (547) were unanimously labelled "geospatial" across all five LLM runs. We manually verified 561 true weak positives and identified 67 false weak positives.

We also randomly sampled and manually verified 572 weak negatives, identifying 7 false weak negatives. Thus, the gold dataset consists of 568 positives and 632 negatives (total size of $1,200$).

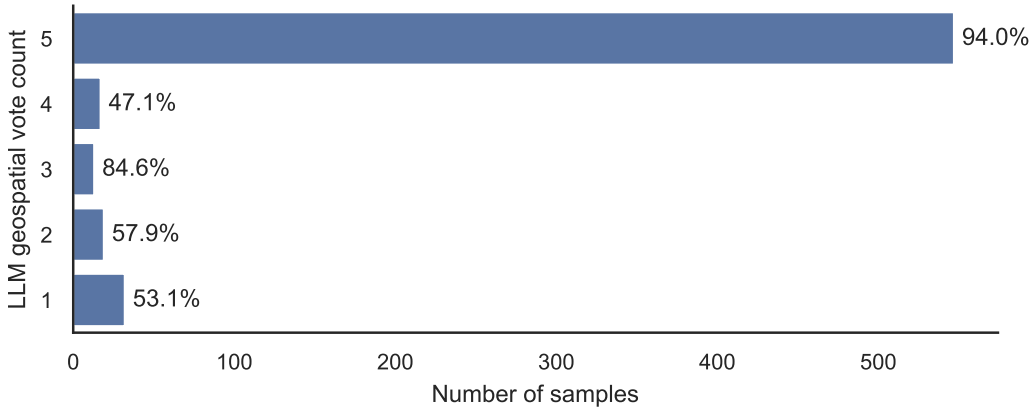


Figure 2 Human verification of weak geospatial labels across LLM vote counts. Bars indicate the number of queries receiving 1-5 spatial votes across repeated LLM runs; percentages denote the proportion confirmed as geospatial after manual verification.

Figure 2 shows human agreement with the weak LLM labels as a function of the number of "geospatial" votes (between 1 and 5). The distribution is heavily skewed toward unanimous predictions, suggesting that most weakly-positive queries are unambiguous for the Llama-3.1 model under the prompt and the definition adopted. The intermediate bins (2-4 votes) contain few queries, so their confirmation rates should not be read as a monotonic trend. Unsurprisingly, the more obvious geospatial queries, such as "what county is springfield, or" (*613049*) or "which region is egypt located" (*1018537*) typically received all 5 positive votes; more borderline cases, such as "biggest house in the world price" (*53174*) or "where do people think dolphins live" (*971381*) produced inconsistent votes across runs, with the model split between geospatial and non-geospatial judgements.

To assess inter-annotator agreement, the second and third authors independently re-annotated 200 samples drawn from the gold set. The lead, second and third authors all agreed on 179 of 200 samples (89.5%); of those, 107 were negative and 72 positive. Pairwise Cohen's κ was 0.88 (lead vs second author), 0.84 (lead vs third), and 0.84 (second vs third), indicating "almost perfect" agreement [26]. As expected, disagreements concentrated on borderline cases. For example, only the lead author read "va" as Virginia in the query "va tax contact" (*535301*); "cricket wireless address" (*112731*) was labelled as geospatial by the lead and second authors but not the third; "biggest house in the world price" (*53174*) was labelled as geospatial by the lead and third authors, but not the second.

4.2 SetFit classifier

When trained on just 200 samples (105 negative and 95 positive) and tested on 800 samples (421 negative and 379 positive) of the gold dataset, the SetFit binary classifier achieves an accuracy of 0.934 (95% bootstrap CI $0.915 - 0.950$) and an F_1 score of 0.930 (95% bootstrap CI $0.909 - 0.947$), producing 27 false negatives and 26 false positives. The majority of misclassified samples are borderline geospatial and can be argued both ways; for example, "biggest snakes in world" (FP, *53491*), "when is the good time to see northern lights" (FP,

952893), "address for expresspcb" (FN, *11492*), or "how wide is the base of the great wall" (FN, *1175805*). This result is particularly notable given the small size of the training set and near-equal class distribution in both training and test sets, making F_1 a meaningful performance indicator rather than a result of class imbalance.

We further verify the classifier's robustness by sense-checking its behaviour on made-up, semantically complex queries and short phrases, a subset of which is demonstrated in Table 1.

Table 1 Some examples of correctly classified, semantically complex made-up queries and phrases (*not* part of MS MARCO) to sense-check the trained classifier's behaviour.

Correctly classified as geospatial	Correctly classified as non-geospatial
nearest hospital	near impossible
countries bordering ukraine	borderline invisible
distance from paris to berlin	keep your distance
restaurants along the m1	stop it m8
where should i live	where can i escape mentally
flood risk in this area	not my area of expertise
is it a long flight london to kaunas	a long way to go in my career

After retraining the classifier on the full gold dataset consisting of 632 negative and 568 positive cases, we run it on the entire MS MARCO dataset consisting of 1,010,916 queries, and identify 181,827 spatial queries, or 18.0% of the entire dataset.

An analysis of the most frequent first words in each class (see Table 2) reveals that *what* dominates both geospatial and non-geospatial queries (29.6% and 36.1% respectively), *how* appears in the top-3 for both (11.6% and 17.9%), while *where* ranks second in geospatial queries (15.8%) but only 14th in non-geospatial queries (0.8%). The overlap between the two vocabularies is striking: 14 of the top-20 first words appear in both lists.

Table 2 Top-20 most frequent first words in geospatial and non-geospatial MS MARCO queries.

Geospatial						Non-geospatial					
#	Word	%	#	Word	%	#	Word	%	#	Word	%
1	what	29.6	11	cost	1.0	1	what	36.1	11	define	1.0
2	where	15.8	12	population	0.9	2	how	17.9	12	definition	1.0
3	how	11.6	13	what's	0.9	3	who	3.7	13	cost	0.9
4	average	3.4	14	does	0.6	4	is	3.0	14	where	0.8
5	when	3.3	15	can	0.5	5	when	2.6	15	do	0.7
6	weather	2.8	16	most	0.5	6	can	2.1	16	the	0.6
7	is	2.5	17	largest	0.5	7	why	1.8	17	are	0.6
8	which	1.9	18	distance	0.5	8	which	1.8	18	meaning	0.5
9	who	1.8	19	temperature	0.5	9	does	1.3	19	what's	0.4
10	why	1.1	20	the	0.4	10	average	1.2	20	causes	0.4

4.3 Grid search for clustering

The top-10 UMAP and HDBSCAN configurations by DBCV, which produced at least 10 clusters, are shown in Table 3.

Table 3 Top-10 UMAP and HDBSCAN hyper-parameter configurations by DBCV score, restricted to those producing at least 10 clusters. *Dims* is UMAP output dimensionality, *Neigh* is UMAP `n_neighbors`, *Min cluster size* and *Min samples* are HDBSCAN's `min_cluster_size` and `min_samples` parameters; *Noise* is the proportion of queries assigned to the noise ('−1') cluster.

Config ID	UMAP		HDBSCAN		DBCV	Noise	Num clusters	Median cluster size
	Dims	Neigh	Min cluster size	Min samples				
93	15	25	200	40	0.358	0.359	102	449.5
79	15	10	100	50	0.341	0.359	182	246
81	15	10	200	40	0.337	0.388	122	476
22	5	25	200	100	0.334	0.389	91	467
21	5	25	200	40	0.333	0.362	105	469
95	15	25	200	200	0.324	0.417	77	553
9	5	10	200	40	0.319	0.392	122	443.5
45	10	10	200	40	0.317	0.374	114	475
23	5	25	200	200	0.315	0.426	79	485
76	15	10	50	25	0.312	0.418	415	119

Table 4 reports the consistency of the top-10 configurations across six repeated runs with different random seeds. Six configurations are stable, with standard deviation for both DBCV and noise fraction below 0.050.

Table 4 Consistency of top-10 configurations across six repeated runs with various random seeds $\{0, 1, 2, 3, 4, 42\}$, reporting minimum, mean, maximum, and standard deviation of the DBCV scores (1=perfect), noise fraction, and number of produced clusters.

Config ID	DBCV				Noise				Num clusters			
	Min	Mean	Max	Std	Min	Mean	Max	Std	Min	Mean	Max	Std
93	0.245	0.364	0.712	0.176	0.000	0.332	0.432	0.165	3	88	109	41.9
45	0.097	0.363	0.811	0.242	0.000	0.245	0.400	0.190	4	77	115	56.3
23	0.279	0.326	0.416	0.049	0.341	0.409	0.464	0.040	64	75	82	6.4
21	0.257	0.310	0.367	0.040	0.358	0.389	0.415	0.027	100	108	114	6.2
81	0.178	0.306	0.358	0.069	0.001	0.244	0.388	0.189	4	77	122	56.6
95	0.227	0.289	0.348	0.044	0.410	0.432	0.470	0.026	72	76	82	3.4
79	0.247	0.282	0.341	0.033	0.344	0.395	0.442	0.038	173	195	208	14.3
76	0.238	0.279	0.312	0.031	0.412	0.421	0.443	0.011	406	421	459	19.4
9	0.092	0.271	0.412	0.109	0.000	0.253	0.406	0.196	4	79	122	58.3
22	0.231	0.269	0.334	0.040	0.389	0.434	0.461	0.033	87	96	101	5.4

Based on the consistency analysis, configurations #21 and #23 are superior; we pick configuration #21 as it achieves a lower mean share of noise (0.389 vs 0.409 in configuration #23) and a higher number of clusters (mean 108 vs 75) despite having a slightly lower mean DBCV (0.310 vs 0.326). For the seed, we pick 42, whose DBCV falls near the configuration's median across runs (we deliberately avoid selecting the highest-DBCV seed to prevent cherry-picking).

4.4 Final clustering

The radial hierarchy chart in Figure 3 illustrates 88 categories (collapsed from 105 initial clusters), grouped into nine broad themes.

Examples of clusters' representative queries, top-20 most frequently occurring terms, and random samples of 10 queries are illustrated in Table 5. Table 6 in the Appendix shows resulting labelled clusters with sizes and representative queries. We make all cluster-level data available on GitHub[15].

[15] `https://github.com/ilyankou/ms-marco-geospatial`/blob/main/interim/cluster_review.md

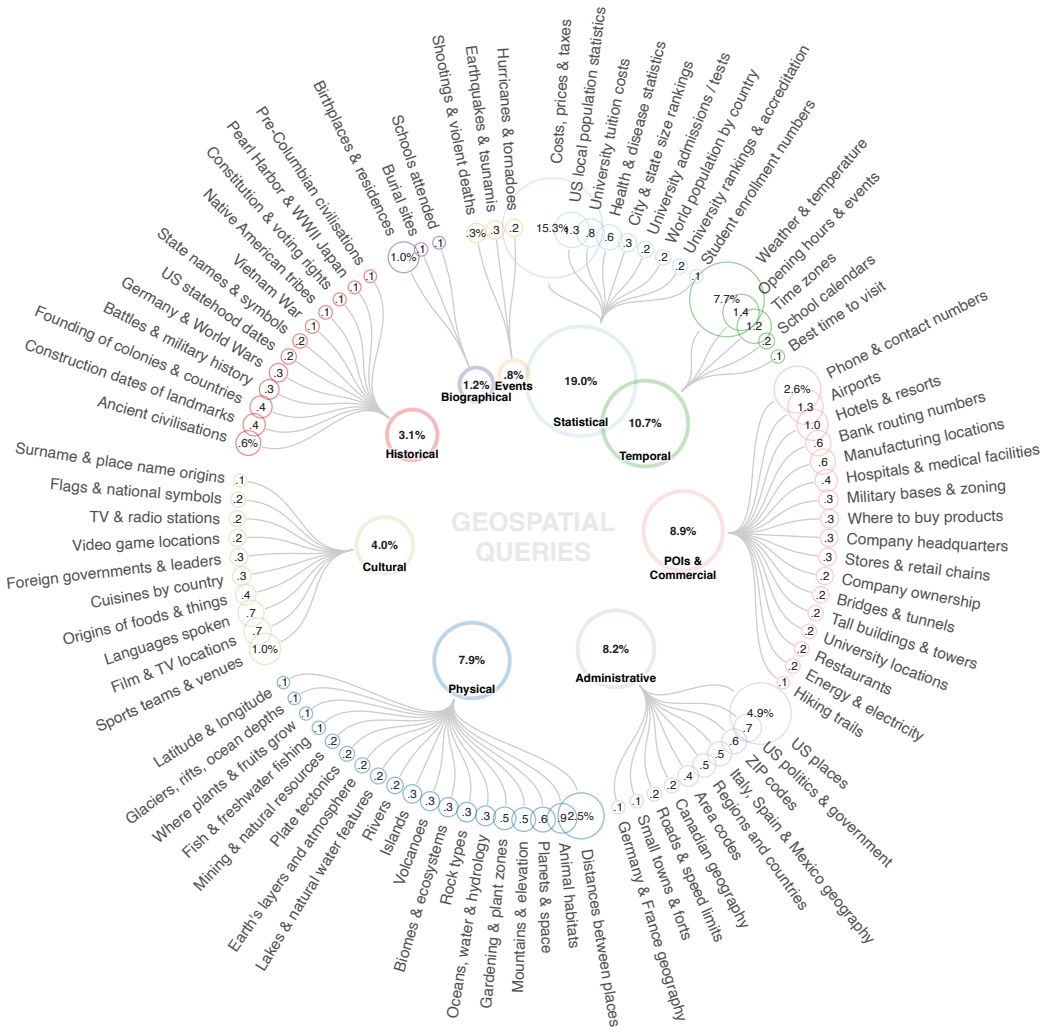


Figure 3 Geospatial query clusters grouped into nine themes. Unclustered (noise) queries representing 36.2% of identified geospatial queries are not shown.

5 Discussion

5.1 Characteristics of the taxonomy

This study identifies 181,827 geospatial queries in MS MARCO, representing 18.0% of the dataset, which is a substantially higher proportion than the 6.17% of queries labelled as *Location* in the original MS MARCO annotations [5]. This discrepancy reflects the breadth of our geospatial definition, which captures implicitly location-anchored queries such as weather conditions, prices, and biomes – categories that may not be caught by keyword-based approaches.

18 of 105 initial clusters (*#55*, and *#88–#104* in Table 6 in the Appendix) largely represent variants of the query "what county is [place] in", differentiated primarily by US state abbreviation. This fragmentation is partly an HDBSCAN artefact as embeddings likely treat state abbreviations, such as "tx" or "ny", as distinguishing features. We group these

Table 5 To derive initial cluster labels, we looked at each cluster's representative query, top-20 most frequent terms, and ten randomly sampled queries.

	Cluster #48	**Cluster #55**
No. samples	344	1015
Representative query	what is the tallest building in the world	what is the county of florida
Top-20 terms	tower, building, pisa, tower pisa, tallest, leaning, leaning tower, tall, tallest building, empire state, floors, building world, empire, state building, world, towers, khalifa, burj, burj khalifa, skyscraper	fl county, fl, county, florida, florida county, beach, palm, beach fl, population, florida located, jacksonville, fl population, county florida, beach florida, located, county palm, springs, orange, county jacksonville, naples
Ten sample queries	largest moving structure in the world · what is the pyramid shaped building on the skyline of san francisco · leaning tower of pisa facts · how many floors is tower one · how many floors does the tallest building in sydney have · leaning tower locale · how many feet is the world's tallest building · how tall is victory tower · how many floors in the empire state building. · where is cornerstone condos located	homes for sale st. cloud fl. · where is dunedin florida · where is naples manor fl · what county is atlantic beach fl in · what county is harbor springs in · where is legoland florida located · what county is ft. lauderdale · homes for sale jacksonville nc · where is citrus florida located · what county is holt fl
Cluster label	Tall buildings & towers	County membership and places in Florida → US places
Theme	POIs & Commercial	Administrative

fragmented clusters into a single *US places* category for Figure 3. HDBSCAN treats 36.2% of identified geospatial queries as noise (not shown in Figure 3), a substantial but unsurprising proportion given the scale and messiness of the real-world dataset. Natural-language web search queries are inherently heterogeneous, and some degree of noise is expected and acceptable in density-based clustering applied at this scale. The *Planets & space* cluster (1,045 largely astronomical queries, which our definition excludes) likely reflects the binary classifier not having been exposed to such queries during training; representing only 0.6% of all geospatial queries, its effect on our analysis is negligible.

We use nine broad themes – Statistical, Temporal, "POIs and Commercial", Administrative, Physical, Cultural, Historical, Biographical, and Events – for organisational convenience rather than as a theoretically grounded hierarchy. We experimented with alternative groupings but found that meaningful structure resides in the leaf categories themselves; any higher-level grouping risks implying a cleaner taxonomy than the data supports. Because we report theme shares over the full set of geospatial queries, including noise, the percentages are conservative; the relative ranking of themes holds only insofar as noise is not concentrated in particular intents, which we do not test here.

The taxonomy reveals that geospatial web search is dominated by transactional and practical lookups rather than by either administrative or physical geography. The Statistical theme alone accounts for 19.0% of geospatial queries and the Temporal theme a further 10.7%. The *Costs, prices & taxes* (15.3%) cluster, with queries such as "what is the average household income of lake oswego" (*1152138*) or "what is currency in brazil" (*736719*), is comparable in size to the Physical (7.9%) and Administrative (8.2%) themes combined. Users

overwhelmingly turn to web search for place-anchored facts about money, time, and weather; questions about landforms, hydrology, or jurisdictional boundaries – traditional GIS territory – form a comparatively small share of expressed information needs.

5.2 Implications for system design

A large share of the taxonomy consists of clusters whose answers are open, stable, and objective. For example, *Distances between places* (e.g., "how far is it from hartford to new london", *231318*) and *ZIP codes* (e.g., "what is burlington,ky zip code", *726469*) are answerable deterministically using geoparsing, spatial databases and a traditional GIS toolkit, including distance calculations and containment operations. Query categories such as *Regions and countries* (e.g., "what is the continent is algeria located", *1151645*), *Bridges & tunnels* (e.g., "where does golden gate bridge connect to", *973130*) or *Volcanoes* (e.g., "what type of volcano is mt.santorini", *915164*) are largely answerable from structured knowledge contained in geographical knowledge graphs. Such types of queries do not require generative or probabilistic systems to be answered correctly. This suggests that a hybrid retrieval architecture of routing objectively answerable geospatial queries to structured geodata sources while reserving generative models for more complex or subjective queries could improve both accuracy and reliability compared to treating all geospatial queries uniformly.

Categories such as *Best time to visit* (e.g., "most affordable times to visit hawaii", *456340*), *University rankings & accreditation* (e.g., "is codarts a good school?", *406730*), and *Origins of foods & things* (e.g., "where does spaghetti bolognese come from", *973905*) are more subjective as their answers depend on personal preferences and the source of information used to answer them. These are the clusters where Anglophone-Western geographic and cultural biases in both the MS MARCO dataset and any downstream system are most likely to surface. Systems trained to answer such queries risk encoding and amplifying the biases of their training data under the guise of objective geospatial information. This risk is sharpest in conversational interfaces, which can deliver an evaluative or unverified answer with the same fluency as a deterministic lookup [20].

Temporal or POI-specific categories such as *Opening hours & events* (e.g., "what day does the garrett pool open", *616654*), *Phone & contact numbers* (e.g., "northampton county tax department pa phone number", *1092161*), or *Restaurants* (e.g., "what restaurant serves both mexican and steak", *1109148*) demand answers that are not only geographically specific but temporally volatile: opening hours, phone numbers, and menus change, and businesses close and relocate much more frequently than coastlines redraw themselves. Unlike physical geography, which changes slowly, institutional information goes out-of-date quickly, making it particularly challenging for static retrieval systems to serve reliably without access to up-to-date data. The stakes are uneven across clusters: an outdated restaurant menu or opening time may cause minor inconvenience, whereas obsolete information about hospital locations or emergency services could have serious consequences. The *Weather & temperature* category (e.g., "what is the weather in nashville tomorrow", *853623*), and the Events theme, accounting for 0.8% of identified geospatial queries and covering *Hurricanes & tornadoes* (e.g., "where there storms last night", *1001188*), *Earthquakes & tsunamis* (e.g., "was there a earthquake in alaska", *541104*), and *Shootings & violent deaths* (e.g., "who is the woman that beat up woman at stop light in houston texas", *1042188*), highlight the relationship between spatial and temporal information needs, which is another argument for hybrid architectures that can handle APIs, news and live data feeds in addition to static geographic knowledge.

The most significant implication concerns emerging natural-language interfaces, including conversational chatbot search. If nearly one in five web search queries is geospatial, these systems require robust geospatial reasoning for a substantial fraction of real user traffic. Existing geographic benchmarks for LLMs have largely focused on factual and location reasoning [6, 28, 37, 30], with comparatively less attention to the transactional, temporal, and subjective queries that dominate our taxonomy. Costs and prices, opening hours, and travel recommendations are examples of query categories where retrieval-augmented and real-time architectures matter most, yet they are largely absent from current evaluation benchmarks. Our taxonomy offers a more representative target for both system design and benchmarking, grounded in what users actually ask rather than what is easiest to test.

5.3 Limitations and future work

MS MARCO was collected prior to 2018 and reflects search behaviour on Bing at that time; the distribution of geospatial query types has likely shifted as search interfaces, user habits, and geospatial services have evolved. The dataset over-represents Anglophone North American users, and this is reflected directly in the taxonomy, where *US places* (4.9%) accounts for vastly more place-specific queries than *Italy, Spain & Mexico geography*, *Canadian geography*, and *Germany and France geography* clusters combined (0.8%). Future work should seek to validate and extend the taxonomy using more geographically diverse query corpora.

The original MS MARCO dataset specifically excludes queries "with navigational and other intents" [5], meaning that a distinct and practically significant category of geospatial information need is not represented in our taxonomy. Our 18.0% is therefore best read as a lower bound, since routing and directions (likely to be a significant geospatial category, even in web search) are excluded by construction.

The corpus reflects web search habits and as such consists of short and typically single-intent queries, which constrains the taxonomy. Longer, multi-intent interactions typical of conversational AI systems, voice assistants, or chatbots fall outside what this analysis captures. In such systems, geospatial intent may be implicit, distributed across dialogue turns, or conditioned on prior context in ways that single-query classification cannot detect. Whether the underlying information needs remain stable across interfaces, or shift as users adapt to conversational systems, is an open empirical question that we aim to address in future work.

Several directions emerge from this work. First, the taxonomy clusters could be linked to existing geospatial datasets and spatial operations (for example, mapping distance queries to routing APIs, or county membership queries to administrative boundary datasets), providing a concrete bridge between user intent and existing GIS infrastructure. Second, the binary geospatial classifier could be extended into a multi-class router that directs queries to category-specific retrieval pipelines. The unclustered "noise" queries, representing over a third of all identified geospatial queries, also deserve further study; repeating the full clustering pipeline on this subset alone may reveal latent structure that the current parameter settings were too coarse to resolve. Finally, the distinction between objective and subjective geospatial queries merits a more formal investigation, as it has direct implications for system design, bias mitigation, and the appropriate use of generative versus deterministic systems.

6 Conclusion

Geospatial web search queries are acts of spatial cognition. When someone types "in which county is north hollywood california" (*394665*) or "cost of living in turkmenistan" (*105311*), they are navigating a conceptually structured space of places, attributes, and relationships.

Our finding that 18.0% of MS MARCO queries are geospatial under a semantically broad definition is evidence that geospatial thinking pervades everyday information behaviour at a scale that toponym-based studies in particular have underestimated.

The derived taxonomy shows that this everyday spatial cognition diverges from what GIS has traditionally modelled. Questions about landforms, hydrology, and physical geography are outnumbered by transactional and practical lookups about costs, weather, and opening hours. This has direct consequences for systems designed to serve such queries.

The taxonomy also surfaces a more fundamental split: queries about distances, ZIP codes, or county membership presuppose that place has determinate answers; queries about the best place to live or the right season to visit presuppose that place is evaluative and culturally inflected. These two modes – geospatial information as fact versus geospatial information as judgement – demand different infrastructure and raise different risks. Distinguishing them is, we argue, a precondition for handling geographic knowledge responsibly.

References

1 Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-51935-3_34.

2 Dimo Angelov. Top2Vec: Distributed Representations of Topics, August 2020. arXiv:2008.09470.

3 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, USA, January 2007. Society for Industrial and Applied Mathematics. URL: https://dl.acm.org/doi/10.5555/1283383.1283494.

4 Ira Assent. Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery*, 2(4):340–350, 2012. doi:10.1002/widm.1062.

5 Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, October 2018. doi:10.48550/arXiv.1611.09268.

6 Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are Large Language Models Geospatially Knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, pages 1–4, New York, NY, USA, December 2023. Association for Computing Machinery. doi:10.1145/3589132.3625625.

7 Ricardo J. G. B. Campello, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1343, 2020. doi:10.1002/widm.1343.

8 Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer. doi:10.1007/978-3-642-37456-2_14.

9 Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. doi:10.1177/001316446002000104.

10 Tim Cresswell. *Place: An Introduction*. John Wiley & Sons, August 2014.

11 Alistair J. Edwardes and Ross S. Purves. A Theoretical Grounding for Semantic Descriptions of Place. In J. Mark Ware and George E. Taylor, editors, *Web and Wireless Geographical Information Systems*, pages 106–120, Berlin, Heidelberg, 2007. Springer. doi:10.1007/978-3-540-76925-5_8.

12 Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, pages 49–56, New York, NY, USA, April 2008. Association for Computing Machinery. doi:10.1145/1367798.1367806.

13 Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models, 2024. doi:10.48550/arXiv.2407.21783.

14 Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022. doi:10.48550/arXiv.2203.05794.

15 Stefan Hahmann and Dirk Burghardt. How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science*, 27(6):1171–1189, June 2013. doi:10.1080/13658816.2012.743664.

16 Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. Place Questions and Human-Generated Answers: A Data Analysis Approach. In *Proceedings of the 22nd AGILE Conference on Geographic Information Science*, pages 3–19, 2019. doi:10.1007/978-3-030-14745-7_1.

17 Ehsan Hamzei, Stephan Winter, and Martin Tomko. Initial Analysis of Simple Where-Questions and Human-Generated Answers. In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *LIPIcs, Volume 142, COSIT 2019*, volume 142, pages 12:1–12:8. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.COSIT.2019.12.

18 Ehsan Hamzei, Stephan Winter, and Martin Tomko. Place facets: a systematic literature review. *Spatial Cognition & Computation*, 20(1):33–81, January 2020. doi:10.1080/13875868.2019.1688332.

19 Andreas Henrich and Volker Luedecke. Characteristics of geographic information needs. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, GIR '07, pages 1–6, New York, NY, USA, November 2007. Association for Computing Machinery. doi:10.1145/1316948.1316950.

20 Ilya Ilyankou, Stefano Cavazzi, and James Haworth. The Scenic Route to Deception: Dark Patterns and Explainability Pitfalls in Conversational Navigation, March 2026. doi:10.48550/arXiv.2603.14586.

21 Ilya Ilyankou, Aldo Lipani, Stefano Cavazzi, Xiaowei Gao, and James Haworth. Do Sentence Transformers Learn Quasi-Geospatial Concepts from General Text? *GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland*, 2024.

22 Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. Quantifying Geospatial in the Common Crawl Corpus. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '24, pages 585–588, New York, NY, USA, November 2024. Association for Computing Machinery. doi:10.1145/3678717.3691286.

23 Rosie Jones, Wei V. Zhang, Benjamin Rey, Pradhuman Jhala, and Eugene Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, March 2008. doi:10.1080/13658810701626186.

24 Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Maria-Aggeliki Pollali, Pierre Maret, and Manolis Koubarakis. The question answering system GeoQA2 and a new benchmark for its evaluation. *International Journal of Applied Earth Observation and Geoinformation*, 134:104203, November 2024. doi:10.1016/j.jag.2024.104203.

25 Werner Kuhn, Ehsan Hamzei, Martin Tomko, Stephan Winter, and Haonan Li. The semantics of place-related questions. *Journal of Spatial Information Science*, 1(23):157–168, 2021. doi:10.5311/JOSIS.2021.23.161.

26 J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

27 Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions. *AGILE: GIScience Series*, 2:1–21, June 2021. doi:10.5194/agile-giss-2-8-2021.

28 Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. GeoLLM: Extracting Geospatial Knowledge from Large Language Models, February 2024. arXiv:2310.06213 [cs]. doi:10.48550/arXiv.2310.06213.

29 Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. doi:10.48550/arXiv.1802.03426.

30 Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '23, pages 85–94, New York, NY, USA, November 2023. Association for Computing Machinery. doi:10.1145/3615886.3627745.

31 Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-Based Clustering Validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, pages 839–847, 2014. doi:10.1137/1.9781611973440.96.

32 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, March 2023. doi:10.48550/arXiv.2210.07316.

33 D. Punjani, K. Singh, A. Both, M. Koubarakis, I. Angelidis, K. Bereta, T. Beris, D. Bilidas, T. Ioannidis, N. Karalis, C. Lange, D. Pantazi, C. Papaloukas, and G. Stamoulis. Template-Based Question Answering over Linked Geospatial Data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, pages 1–10, Seattle WA USA, November 2018. ACM. doi:10.1145/3281354.3281362.

34 Ross S. Purves, Stephan Winter, and Werner Kuhn. Places in Information Science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182, 2019. doi:10.1002/asi.24194.

35 Mansi Radke, Prarthana Das, Kristin Stock, and Christopher B. Jones. Detecting the Geospatialness of Prepositions from Natural Language Text. In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *14th International Conference on Spatial Information Theory (COSIT 2019)*, volume 142 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:8, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.COSIT.2019.11.

36 Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1410.

37 Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. GPT4GEO: How a Language Model Sees the World's Geography, May 2023. arXiv:2306.00020 [cs]. doi:10.48550/arXiv.2306.00020.

38 Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval. 27th Annual International ACM SIGIR Conference*, 2004.

39 Benny Shanon. Answers to where-questions. *Discourse Processes*, 6(4):319–352, 1983. doi:10.1080/01638538309544571.

40 Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.135.

41 Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient Few-Shot Learning Without Prompts, September 2022. doi:10.48550/arXiv.2209.11055.

42 Haiqi Xu, Ehsan Hamzei, Enkhbold Nyamsuren, Han Kruiger, Stephan Winter, Martin Tomko, and Simon Scheider. Extracting interrogative intents and concepts from geo-analytic questions. *AGILE: GIScience Series*, 1:1–21, July 2020. doi:10.5194/agile-giss-1-23-2020.

A LLM prompt for weak labelling

Listing 1 Prompt for Llama-3.1 to label 5,000 sampled queries as geospatial/non-geospatial.

```
Classify the search query as geospatial (true) or not (false).

A query is geospatial if it requires qualitative or quantitative geographic
knowledge of Earth-bound features to be answered. This is usually the case if the
    query involves:
- A geographic entity (named place on Earth: city, country, river, POI, address)
- A geographic concept (place type: city, lake, mountain, park, building)
- A spatial relation (near, within, north of, between, borders, crosses, distance)

Non-geospatial: anatomical, microscopic, astronomical, fictional, or abstract
'where' questions; queries needing no geographic knowledge.

Output only: true or false.

Examples:
  * How far is Brighton from London -> true
  * what does square from greece mean -> false
  * Capital of France -> true
  * What does 'Dutch courage' mean -> false
  * Population of Piraeus -> true
  * Where is the epimysium found -> false
  * Restaurants near Hyde Park -> true
  * Where did the name Missouri come from -> false
  * Where does Paris Hilton live -> true
  * What is a river -> false
  * What city is Ebright Azimuth in -> true
  * is the water underneath a hurricane calm -> false
  * how many state are in the us -> true
  * Where is the SSID -> false
  * where are new franchises needed -> true
  * which side is the us flag put -> false
  * what languages are spoken in israel -> true
  * Where is Hogwarts -> false
  * most western point in portugal -> true
  * What is Greek yoghurt -> false

Query: <SEARCH QUERY>
Answer:
```

B Clusters with representative queries

Table 6 Resulting clusters with final labels and representative queries.

#	Size	Label	Representative query *(query ID)*
-1	65823	Unclustered (HDBSCAN noise)	what is the url for the state bank of the lakes *(852466)*
0	696	Area codes	area code numbers united states *(26408)*
1	1054	ZIP codes	what's the zip code *(933999)*
2	13966	Weather & temperature	what's the temperature outdoors *(933263)*
3	27761	Costs, prices & taxes	what is the current minimum wage for ny *(1151495)*
4	526	Earthquakes & tsunamis	what is been the biggest earthquake *(722978)*
5	1344	US politics & government	how many representatives for each state *(294437)*
6	596	Military bases & zoning	what military base is the largest in the us *(878154)*
7	298	State names & symbols	when was the state nickname chosen *(962316)*
8	1296	Film & TV locations	where was where the heart is filmed *(1004279)*
9	217	Schools attended	what college did trump attend *(597060)*
10	1744	Birthplaces & residences	where was mlk born *(1002963)*
11	505	City & state size rankings	what is the largest city in the US by area *(827081)*
12	818	Hospitals & medical facilities	university of iowa hospitals and clinics patient *(532623)*
13	316	Canadian geography	toronto is in what province *(522784)*
14	441	Hurricanes & tornadoes	where does a hurricane occur *(972347)*
15	1406	University tuition costs	usu tuition cost per semester *(534985)*
16	1096	Bank routing numbers	us bank routing numbers *(533522)*
17	4641	Phone & contact numbers	state street service desk number *(503076)*
18	246	Best time to visit	when is the best time to visit the bahamas *(952523)*
19	287	Roads & speed limits	types of roads and their speed limits *(529594)*
20	2228	Time zones	what time zone is in *(905691)*
21	2571	Opening hours & events	what time does shoe carnival close *(904553)*
22	201	Hiking trails	how long is the trail to hike the narrows *(265295)*
23	300	Energy & electricity	what energy source is used most by the usa *(657321)*
24	206	Latitude & longitude	latitude is measured from where *(438015)*
25	301	Restaurants	what restaurant are open *(891272)*
26	307	University locations	which state is harvard university located *(1019701)*
27	449	Company ownership	who owns the ritz carlton chain *(1045915)*
28	1801	Hotels & resorts	where is the resort that has the rooms over the water *(997370)*
29	476	Company headquarters	where is headquarters located *(984270)*
30	1797	Sports teams & venues	where is the college football playoff *(995090)*
31	375	TV & radio stations	what networks stream live tv *(881802)*
32	367	School calendars	what day does school start in florida *(616616)*
33	2368	Airports	what is closest airport *(731434)*
34	4494	Distances between places	how far is it between two cities *(231276)*
35	1082	Manufacturing locations	where does ford manufacture *(973047)*
36	469	Stores & retail chains	how many shops does walmart have *(295729)*
37	526	Where to buy products	Where Can I Buy Cruex *(8025)*
38	249	Burial sites	name of where dead bodies are buried *(461873)*
39	562	Shootings & violent deaths	how many deaths from the shooting in fl *(1097258)*
40	321	University rankings & accreditation	top ranked universities in us *(522650)*
41	247	Student enrollment numbers	how many students does osu have *(297340)*
42	391	University admissions / tests	average gpa acceptance at unr *(36767)*
43	250	Small towns & forts	what part of maine is portland in *(884763)*
44	237	Surname & place name origins	where does the name ireland come from *(974631)*
45	533	Cuisines by country	what are the names of two traditional foods in argentina *(571955)*
46	661	Origins of foods & things	where do chickens come from originally *(1141679)*
47	390	Bridges & tunnels	longest bridge in *(1174000)*
48	344	Tall buildings & towers	what is the tallest building in the world *(849724)*
49	526	Islands	what island is off the map *(865244)*

Table 6 (continued).

#	Size	Label	Representative query *(query ID)*
50	368	Lakes & natural water features	what is the largest freshwater lake in the united states *(827162)*
51	255	Native American tribes	where native american came from *(1000891)*
52	214	Pre-Columbian civilisations	what ancient civilizations were in mexico *(553539)*
53	396	Rivers	which river is the longest *(1139538)*
54	948	Mountains & elevation	which mountain is the tallest *(1013640)*
55	1015	US places	what is the county of florida *(813209)*
56	480	Foreign governments & leaders	what type of government does mexico have currently *(912604)*
57	328	Flags & national symbols	what do the colors of the flag say about the nation *(625172)*
58	440	Video game locations	where do i find quarried stone skyrim *(970929)*
59	1046	Ancient civilisations	what is the ancient name of the area *(804856)*
60	1080	Health & disease statistics	how many sick people in the united states *(295798)*
61	345	World population by country	what the world population *(903847)*
62	615	Oceans, water & hydrology	most of the water on earth is found where *(458406)*
63	210	Glaciers, rifts, ocean depths	what is the deepest part of the ocean *(814391)*
64	2403	US local population statistics	what is the population of wv *(1034203)*
65	811	Construction dates of landmarks	what year was the statue of liberty constructed *(928604)*
66	291	Mining & natural resources	where is silver ore found *(992908)*
67	245	Fish & freshwater fishing	what fish are in lake nottely *(1117541)*
68	271	Vietnam War	when did the us go to vietnam war *(942467)*
69	1182	Languages spoken	what languages are spoken *(872702)*
70	909	Regions and countries	asia where is it located *(27347)*
71	314	Plate tectonics	when oceanic crust and continental crust meet at the plate boundary *(954075)*
72	1045	Planets & space	what is the nearest planet to earth *(836079)*
73	330	Earth's layers and atmosphere	what layer of the earth is on the surface *(872996)*
74	561	Volcanoes	what type of eruption formed the volcano in this photograph *(912074)*
75	598	Rock types	granite is what type of rock *(196556)*
76	239	Germany & France geography	germanys geographic location *(194599)*
77	979	Italy, Spain & Mexico geography	what is italy's location *(761674)*
78	221	Pearl Harbor & WWII Japan	when did the japanese attack pearl harbor *(941810)*
79	236	Constitution & voting rights	how many states voted to ratify the constitution *(296971)*
80	236	Where plants & fruits grow	where do pomegranates grow *(971421)*
81	842	Gardening & plant zones	what planting zone am i *(887313)*
82	579	Biomes & ecosystems	place where desert are found and their plants *(475668)*
83	1594	Animal habitats	where are most of the animals located *(965819)*
84	794	Founding of colonies & countries	dates that colonies were founded quizlet *(115607)*
85	334	US statehood dates	what year did states become states *(926788)*
86	612	Battles & military history	what was the last major battle of the civil war *(921122)*
87	514	Germany & World Wars	what did the construction of the berlin wall do to germany *(619964)*
88	433	US places	in what county is washington wi *(394062)*
89	262	US places	what county in minneapolis mn *(602024)*
90	258	US places	lansing mi is in what county *(435687)*
91	1241	US places	austin texas is in what county *(29756)*
92	474	US places	what county richmond va *(615048)*
93	205	US places	what county is atlanta *(602628)*
94	258	US places	what county is kansas city mo in *(607886)*
95	456	US places	what county is new york ny *(610266)*
96	295	US places	what county is cambridge ma *(603814)*
97	467	US places	in what county is columbus, oh *(393920)*
98	382	US places	what county is chicago in illinois *(604215)*
99	555	US places	what county in lebanon pa in *(602014)*
100	213	US places	what county is jersey city nj *(607788)*
101	750	US places	what county is farmville nc in *(605858)*
102	342	US places	what county nashville tn in *(615017)*
103	869	US places	valencia is in what county in ca *(1089047)*
104	381	US places	what county portland in *(615046)*

Georeferencing Non-Gazetteered Place Names Using Biological Specimen Records

Aneesha Fernando
School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Surangika Ranathunga
School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Kristin Stock
School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Raj Prasanna
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand

Christopher B. Jones
School of Computer Science and Informatics, Cardiff University, UK

Abstract

Biological specimen records collected by natural history institutions constitute a rich source of temporal geographic knowledge, capturing biodiversity information about regional landscapes as they were recorded at different times. Using digitised data from the Allan Herbarium (New Zealand), this study identifies place names in these specimen locality descriptions that are absent from current gazetteers; we refer to these as non-gazetteer place names (NGPs). These place names are typically historical, vernacular, or colloquial and were used as landmarks to describe a specimen's location at the time of collection. We then investigate the problem of georeferencing the NGPs using only the limited information available in the specimen records. To resolve this, we leverage repeated occurrences of the same place name across specimen records with different specimen locations and spatial relation terms, extracting and inverting these relations to derive constraints on NGP locations. This approach is instantiated within deterministic, probabilistic, and LLM-based methods, enabling a comparative analysis of their strengths and limitations for text-based spatial inference. On a pseudo-NGP benchmark, probabilistic inference achieves the highest accuracy (median error 1.43 km; A@1 km 36%), while the LLM yields competitive but less precise estimates (median error 1.80 km; A@1 km 31%), indicating that, despite advances in LLMs, traditional modelling remains advantageous when high spatial precision is required.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Natural language processing; Mathematics of computing → Probabilistic inference problems

Keywords and phrases Non-gazetteered Place Names, Gazetteers, Toponyms, Georeferencing, Spatial Relation Modelling, Large Language Models

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.11

Supplementary Material *Dataset*: `https://doi.org/10.6084/m9.figshare.32397048`

Funding This work was supported by the Ministry of Business Innovation and Employment Smart Ideas Fund (grant number MAUX2104), New Zealand.

1 Introduction

The content and metadata associated with biological specimen collection records constitute a rich and underutilised source of geographic information. These records usually include locality descriptions that are routinely recorded digitally by natural history institutions, including museums and herbaria [6, 28]. A typical locality description specifies the location where a

© Aneesha Fernando, Surangika Ranathunga, Kristin Stock, Raj Prasanna, and Christopher B. Jones; licensed under Creative Commons License CC-BY 4.0

biological specimen was collected and usually contains at least one place name, often several, along with spatial relation terms that describe the specimen's position relative to those places. Such records span more than a century and capture geographic information as it was recorded at the time of collection. Consequently, some of the place names recorded in these descriptions are not represented in current gazetteers and can be classified as non-gazetteer place names (NGPs). We define NGPs as place names that are entirely absent from existing gazetteers, including alternative names and other variants, and therefore have no associated spatial footprint in those gazetteers. Their absence may stem from temporal changes in toponymy, the use of informal or locally recognised names, or incomplete coverage of certain regions and feature types in existing gazetteers [14]. Given that millions of such records are available worldwide, these descriptions represent a significant and largely untapped source of geographic knowledge. Incorporating these place names, together with their inferred spatial footprints, into digital gazetteers has the potential to improve spatial reference coverage and support more robust place name resolution, not only for specimen georeferencing but also for broader geographical information retrieval applications and analyses.

In this research, we focus on georeferencing NGPs using information available within biological specimen records. The information contained in biological specimen records includes free-text locality descriptions indicating where specimens were collected, coordinates of collection locations, habitat descriptions, and higher-level administrative units such as states or provinces.[1] These elements are used jointly for both georeferencing and disambiguation of place names in locality descriptions, with specimen coordinates serving as important spatial anchors that constrain the interpretation of relative spatial relations and reduce uncertainty in the inferred locations of unknown place names.

Biological specimen collection practices further provide a source of spatial redundancy that can prove valuable. During fieldwork, collectors typically traverse a specific route and record multiple specimens collected within a limited geographic area. As a result, many specimens share common geographic features and place names across their locality descriptions. In modern collection workflows, specimen coordinates are often recorded in the field and subsequently validated by georeferencers. In historical datasets, however, coordinates may also have been assigned retrospectively from locality descriptions, and as a result, some uncertainty may be present in the coordinates. Moreover, a substantial proportion of records, particularly those collected prior to the widespread adoption of GPS, remain ungeoreferenced, and existing georeferences vary in quality [21]. To address this gap, a range of georeferencing methods has been developed, from time-intensive manual interpretation using GIS tools and maps [28] to automated approaches based on rule-based techniques [26], machine learning [22], and modern Large Language Models (LLMs) [10]. Building on this context, our methodology leverages digitised specimen records and exploits the repeated occurrence of the same NPG across multiple locality descriptions as a source of spatial evidence. When an unknown place name appears in several descriptions, the spatial constraints expressed in each record are combined to infer its most plausible location.

Table 1 illustrates this pattern of repeated place name occurrence across multiple locality descriptions. The NGP *Cabstand*[2] appears in four independent locality descriptions, each linked to a distinct specimen and to the coordinates of the corresponding specimen location.

[1] Example Allan Herbarium specimen record: `https://scd.landcareresearch.co.nz/Specimen/CHR%20508209`.

[2] The publication "Banks Peninsula Conservation Walks" [9], issued by the Department of Conservation, New Zealand, refers to "Cabstand" as the name of a junction on Summit Road in Banks Peninsula.

Each description employs relative spatial relation terms such as "north of," "south of," or "southeast of" to describe the specimen's location relative to "Cabstand," thereby imposing qualitative constraints on its location. Taken together, these multiple references impose complementary spatial constraints that can be exploited to infer the location of the NGP by inverting (reversing) the spatial relations. The associated specimen coordinates act as spatial anchors, bounding the plausible region in which Cabstand must lie. By integrating reversed relative direction terms with the spatial distribution of specimen locations, it becomes possible to approximate the location of "Cabstand" even in the absence of an explicit gazetteer entry.

To operationalise this idea, we investigate three complementary georeferencing methodologies. The first approach is a deterministic method that draws on the formal search space model of Chen et al. [7] and constructs an Approximate Location Region (ALR) for an NGP by modelling the spatial relationship constraints expressed in locality descriptions relative to specimen coordinates. When the same place name appears across multiple descriptions, the intersection of their respective constraint regions progressively reduces the ALR, yielding increasingly precise candidate locations. Second, we adopt a probabilistic framework in which spatial relations are represented as likelihood functions conditioned on specimen coordinates and referenced place names. Multiple occurrences are aggregated to estimate the most likely location from the resulting spatial probability distribution, replacing crisp geometric boundaries with a continuous representation. Third, we explore an LLM-based approach by providing locality descriptions and specimen coordinates as inputs in the prompt and tasking the LLM with inferring NGP locations from combined textual and numeric cues. This allows us to assess the extent to which LLMs can integrate spatial language and coordinate information for georeferencing.

Among these three approaches, the probabilistic method yields the most precise results, although the other approaches also perform competitively. To our knowledge, this study represents the first systematic investigation of using biological specimen records as a primary data source for extracting and georeferencing NGPs. The methodologies and findings presented here demonstrate the feasibility and value of exploiting specimen-derived spatial information for enriching digital gazetteers.

The contributions of this work are as follows:

(1) We demonstrate that biological specimen records contain non-gazetteered place names, highlighting these records as an underutilised source of geographic information over time.

(2) We propose a data-driven strategy to georeference non-gazetteered place names using biological specimen records by exploiting multiple occurrences of the same place name across different specimen locality descriptions to derive spatial constraints on unknown locations.

(3) We provide a systematic comparison of deterministic, probabilistic, and LLM-based approaches to text-based spatial modelling and inference of the locations of non-gazetteered place names, evaluating their strengths and limitations.

(4) We construct a new benchmark dataset for evaluating methods for georeferencing place names derived from biological specimen records, which contain rich spatial knowledge.

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 details the methodology. Section 4 presents the results, and Section 5 concludes the paper with a discussion of limitations and future research directions.

Table 1 Example illustrating how multiple specimen locality descriptions refer to the same non-gazetteered place name ("Cabstand"), each accompanied by distinct coordinates and expressed through varying spatial relation terms (e.g., near, north of, south of, southeast of) that describe the NGP's position relative to the specimen location.

Place name	Locality description	Latitude	Longitude
Cabstand	Banks Peninsula, Brocheries Road near Cabstand	-43.8030	173.0210
	Banks Peninsula, northeast of Akaroa, north of Cabstand, Summit Road	-43.7949	173.0210
	Banks Peninsula, south of Cabstand	-43.8093	173.0172
	Banks Peninsula, Long Bay Road southeast of Cabstand	-43.8039	173.0259

2 Related Work

This section reviews research relevant to the georeferencing of place names absent from gazetteers. It first outlines general approaches to georeferencing place names from textual descriptions, which provide the foundation for the research problem, and then examines studies addressing non-gazetteered place names, clarifying how this problem differs from generic georeferencing and highlighting the existing gap in the literature.

Automated georeferencing links textual descriptions or other location-referencing information to geographic coordinates. While some methods focus on georeferencing entire textual documents, such as Wikipedia articles, news reports, or social media posts [17, 13], others focus specifically on georeferencing the place names mentioned within text descriptions. The latter typically involves two main steps: identifying place names (toponym recognition) and determining their correct locations (toponym resolution or geocoding) [19, 27]. In many applications, both steps are closely tied to gazetteers: recognition benefits from name lists, and resolution selects among candidate gazetteer entries.

Early georeferencing approaches treated toponyms as gazetteer lookups with heuristic disambiguation, favouring populous or administratively prominent places [3, 4]. Unsupervised models improved this by selecting spatially consistent candidate sets using only gazetteer data [16]. However, a fundamental limitation of gazetteer-dependent methods is their incomplete coverage of vernacular, colloquial, and fine-grained place names [14, 1, 24]. As Wang et al. [27] observe, many geoparsing corpora restrict annotations to city-level and above, excluding fine-grained toponyms such as street names or the names of parks and monuments. These methods also struggle to resolve historical or outdated names that are no longer present in modern geographic databases [12]. Nevertheless, such names frequently persist in literary, archival, and other domain-specific textual sources with a historical dimension [3]. To overcome this, gazetteer-independent approaches such as TopoCluster learn spatial word distributions from large georeferenced corpora (GeoWiki) and resolve toponyms based on contextual geographic overlap [8].

Deep learning approaches replaced hand-crafted heuristics with learned contextual representations. For toponym recognition, neural sequence models (e.g., BiLSTM-CRF and CNN–RNN hybrids) capture informal spellings and noisy text, outperforming rule-based and gazetteer-only systems [2, 31, 20]. For toponym resolution, neural models either predict continuous coordinates from contextual embeddings, treating the earth as a regression surface or frame the task as tile classification or shared embedding learning, enabling similarity-based

grounding beyond strict gazetteer lookup [5, 11]. However, these methods rely on large-scale training corpora and still presuppose that targets resemble places represented in their training distributions.

LLMs are beginning to reshape both toponym recognition and resolution by injecting stronger, more explicit geographic knowledge into language representations. Geospatially grounded transformers such as GeoLM [18] and GeoReasoner [29] augment standard pretraining with coordinate-based spatial embeddings and contrastive alignment between textual and geospatial contexts, then fine-tune (or zero-shot apply) these models to sequence-tagging for toponym recognition and toponym linking against gazetteers, achieving state-of-the-art or near state-of-the-art performance on multiple benchmarks. For large-scale resolution, lightweight open-source LLMs (e.g., fine-tuned Mistral, Llama-2, Baichuan2) can be trained to infer an unambiguous type-level referent (city/state/country) from context and then delegate to geocoders such as GeoNames or Nominatim, substantially outperforming previous neural and voting-based resolvers at the global scale [15].

In our study, we target a stricter case of place names that are not found in gazetteers at all, not even as alternate spellings, variants, or official forms. In this setting, resolution cannot fall back on candidate resolution from GeoNames, OpenStreetMap, or similar resources and must infer location from other evidence. To date, relatively few studies have directly targeted the resolution of such truly non-gazetteered place names. A closely related line of work is the Bhugol framework proposed by Sharma et al. [23], which addresses non-gazetteered place names by exploiting external regularities rather than relying on an explicit gazetteer. Bhugol estimates coordinates by clustering either gazetteer place names sharing frequent morphological affixes (via HDBSCAN) or gazetteer place names that co-occur with the target name in a news corpus (via DBSCAN), and assigning the centroid of the most representative spatial cluster as the inferred location; an integrated variant combines both signals. Among existing approaches, this methodology is the closest to our work in its explicit focus on non-gazetteered place names. However, it is strongly region-dependent and relies on recurring naming conventions and extensive corpus evidence, which limits its portability across geographies and languages. Moreover, even its best-performing variant, Bhugol-LS, reports mean distance errors often exceeding 100 km, which may be acceptable for coarse-resolution tasks but falls short in applications requiring fine-grained spatial accuracy.

Chen et al. [7] propose a graph-based approach to georeferencing place names absent from gazetteers by exploiting spatial relationships extracted from natural language, a method that has informed our work. Their framework uses place graphs from collective manually created descriptions, with nodes representing locations and edges encoding qualitative spatial relation terms, enabling inference of a target location by linking a non-gazetteered place name (locatum) to surrounding gazetteered landmarks (relata). However, their approach assumes that the relatum can be reliably identified and disambiguated, and that the locatum corresponds to a formal or alternative gazetteer entry. In contrast, the locality descriptions in our data refer to locations that lack any representation, formal or variant in existing gazetteers or VGI resources. Moreover, our records provide no external contextual cues regarding the prominence or granularity of either locatum or relatum, making it infeasible to construct the probabilistic search space required by their pipeline. Relatedly Yousaf and Wolter [30] propose a reasoning-based model that interprets natural language place descriptions by representing named and unnamed spatial entities as relational constraints and resolving them through spatio-ontological inference. While their approach aligns with our emphasis on spatial relation terms, it ultimately grounds interpretations by matching inferred entities to OpenStreetMap features, whereas our work targets place names that lack any gazetteer representation.

Taken together, the literature suggests a gap in both (i) using specimen records as a source for identifying place names that are absent from gazetteers and (ii) developing methods to georeference place names that are neither gazetteered nor supported by rich corpus-derived evidence. This motivates our effort to extract and georeference non-gazetteered place names using biological specimen records as the source.

3 Methodology

3.1 Locality description corpus

For our experiments, we use biological specimen records from the Allan Herbarium (CHR)[3] dataset of Manaaki Whenua – Landcare Research (MWLR), New Zealand. We retain only records that include both specimen coordinates and a free-text locality description. As described earlier, specimen datasets often contain repeated locality descriptions as multiple specimens may be collected at the same site. Because our objective is to use distinct locality descriptions to constrain an NGP's position, we removed duplicate locality descriptions from the original MWLR extract. We also noted the presence of records that share identical coordinate pairs but differ slightly in wording (e.g., *"Campbell Island, near* ***footpath towards*** *Met Service* ***hostel****"* vs. *"Campbell Island, near* ***concrete pad on*** *Met Service* ***base****"*). To avoid redundant spatial evidence, we removed records with duplicate coordinate pairs, keeping only the first. While this filtering reduces redundancy, it may also discard useful linguistic variation, as records with the same coordinates can contain complementary contextual clues.

Next, we extracted place names from locality descriptions using the *en_core_web_trf* Named Entity Recognition (NER) model in spaCy[4], fine-tuned on a subset of 50 manually annotated locality descriptions from our data. We then retained only records in which at least one extracted place name participates in an explicit spatial relationship (e.g., "18 km west of Ashburton"). Descriptions that merely list place names without explicitly expressing a spatial relation (e.g., "Green Hill, Upper Poulter Valley, Waimakariri Region") were excluded, although in some cases such listings may implicitly encode hierarchical relationships.

The initial list of relation terms was developed from common spatial expressions observed during manual inspection of locality descriptions and was subsequently refined iteratively by checking unmatched and incorrectly matched examples. The matching patterns followed a general structure of the form `<spatial relation term> <place name>`. When a locality description contained multiple such expressions, all matching relations were extracted. A detailed discussion of the spatial relation terms identified in the dataset is provided in Section 3.3.1.

After processing, the dataset comprised 13,746 locality descriptions, grouped by extracted place names, together with their associated spatial relations and the coordinates of each locality description.

3.2 Identifying non-gazetteered place names

To demonstrate the presence of NGPs in biological specimen records, we used a two-stage identification procedure: (i) automated screening to construct a candidate set, and (ii) targeted manual verification to assess its reliability. This procedure was applied to the place names extracted from the above-mentioned dataset of 13,746 locality descriptions. Each

[3] https://www.landcareresearch.co.nz/tools-and-resources/collections/allan-herbarium
[4] https://spacy.io/

extracted place name was checked against the New Zealand Geographic Board (NZGB)[5] Gazeteer, GeoNames[6], and OpenStreetMap (OSM)[7] to determine whether it is already gazetteered.

In the automated stage, we applied sequential filtering using these three gazetteers. First, each candidate was compared against the downloaded NZGB Gazetteer export and matching names were removed. Second, the remaining candidates were cross-checked against the GeoNames export for New Zealand to identify additional gazetteered names not present in the NZGB dataset. Third, unresolved candidates were queried against the OSM Nominatim API, and any strings that could be matched to an OSM feature were excluded. After these steps, 520 candidates were not found in any of the three gazetteers and were retained as candidate NGPs. To reduce sensitivity to minor orthographic variation in the free-text locality field (e.g., misspellings, spacing differences, or abbreviations), we employed fuzzy string matching rather than exact matching, using a similarity threshold of 80% with the RapidFuzz library[8].

Despite this multi-source filtering, the resulting set still required manual verification because specimen locality descriptions are authored as free text and are not standardised. To quantify the extent of these issues, we manually inspected a random sample of 50 of the 520 automatically identified candidates. Of these, 30 were confirmed as true NGPs.

The observed false positives largely fell into five categories:

- **Alternative or informal naming** (a real place, but referenced using a non-standard variant): e.g., Bowscale Lake for Bowscale Tarn; Hutt Road for Hutt Motorway.
- **Generic feature terms misidentified as place names by spaCy** (common nouns treated as named entities): e.g., Airport in "Kaitaia, Quarry Road, near Airport"; BNZ bank in "Waikanae, marae grounds behind BNZ bank".
- **Misspellings** (orthographic noise that persisted despite fuzzy matching): e.g., Mt Rockfort for Mt Rochfort; Baffalo beach for Buffalo Beach.
- **Abbreviations** (shortened forms not captured by gazetteer matching): e.g., big r. for Big River; Cobb V. for Cobb Valley.
- **Compound or delimited forms** (multi-name strings that complicate recognition and matching): e.g., "along the Slaughter/Long Burn"; "near Belltown/Manunui hut".

These error modes highlight that identifying truly non-gazetteered names is not simply a matter of gazetteer lookup. Developing a comprehensive recognition and normalisation pipeline is an important problem in its own right; however, our aim here is to establish that specimen records themselves are a credible source of non-gazetteered names, and the findings in our analysis support this.

Based on the manual audit, we estimated that approximately 60% of the automatically extracted candidates were true NGPs. Applying this precision estimate to the 520 candidates yields an expected ~ 312 true NGP mentions, corresponding to roughly 2% of the 13,746 locality descriptions in our sample (i.e., $520 \times 0.60 \approx 312$; $312/13{,}746 \approx 0.023$). Although this proportion appears small, it is operationally significant at the collection scale: biodiversity infrastructures aggregate millions of specimen records, and even a 2% rate translates into tens of thousands of locality descriptions containing place references that cannot be resolved

[5] https://gazetteer.linz.govt.nz
[6] https://www.geonames.org
[7] https://www.openstreetmap.org
[8] https://pypi.org/project/RapidFuzz

through conventional gazetteer-based workflows. Notably, this proportion is observed in New Zealand, a relatively well-mapped country with a mature national gazetteer and rich VGI sources, suggesting that the prevalence of NGPs is likely to be higher in regions with less complete mapping coverage.

Table 2 presents a representative subset of NGPs from our dataset, categorised according to the most commonly observed (non-exhaustive) NGP types. Many of these NGPs correspond to historical place names that are absent from current gazetteers, yet retain substantial historical value.

Table 2 Examples of manually verified NGPs identified in the dataset, categorised by the most commonly observed (non-exhaustive) types, together with the locality descriptions in which they appear and the corresponding record dates.

Type	Place name	Example occurring locality description	Recorded date
Farm stations	Beltana station	North Canterbury, hills east of Parnassus, above Beltana Station, near radio station	6/01/1969
	Ngatapa Sunworth Station	Gisborne, northwest of Ngatapa Sunworth Station.	24/02/1990
Trig stations	Oropuke trig	Chatham Island, nearest major locality Oropuke trig, Southern Tableland, east of Oropuke trig, at top end of "The Slump", near hut	26/02/1985
	Driscoll Trig	Wairarapa, East Wairarapa Ecological District, Te Wharau near Driscoll Trig	7/08/1992
Infrastructure facilities	Huia Filter Station	Titirangi, Woodlands Road, at Huia Filter Station	16/03/1995
	P. & T. Microwave Station	Vernon Hills, Marlborough, near P. & T. microwave station	24/11/1970
Community facilities	Bryant Home	South shore of Raglan Harbour, near Bryant Home	23/10/1967
	Manaroa School	Forest remnant beside Manaroa school near shore, Pelorus Sound, Marlborough	12/11/1971
Recreation places	Christian Youth Camp	Whakamaru, near Christian Youth Camp	25/02/1983
	Catleys (camp site)	Kauaeranga Valley, Thames opposite Catleys	26/11/1973
Hospitality venues	Cave Rock Tearooms	Sumner near Christchurch, Canterbury, near Cave Rock Tearooms	28/10/1971
	Delta Lodge	Near Delta Lodge, Orongorongo Vly.	25/06/1974

3.3 Georeferencing non-gazetteered place names

We present three methods for georeferencing NGPs using biological specimen records: deterministic, probabilistic, and LLM-based. Each method has distinct strengths and weaknesses. The deterministic and probabilistic methods use structured relation inputs derived from locality descriptions, whereas the LLM-based method operates directly on the unprocessed locality descriptions and associated specimen coordinates. However, as noted in the Introduction, all three are based on the same underlying strategy of exploiting repeated occurrences of the same NGP across multiple locality descriptions as sources of spatial evidence to infer its location.

For example, the locality description *"Main Divide, Coromandel Ra, 1 mile SE of Maumaupaki Trig"* indicates that a specimen was collected one mile southeast of Maumaupaki Trig, a trig station in the Maumaupaki area of New Zealand that is absent from current gazetteers. Because the specimen record includes geographic coordinates (latitude: −36.97612, longitude: 175.5862), this relation can be inverted to infer that the NGP lies one mile northwest of the specimen location.

When multiple descriptions refer to the same NGP using different spatial relations, as illustrated in Table 1, the intersection of these constraints enables a more precise estimate of its location. The following subsections describe how this strategy is operationalised in our study.

3.3.1 Interpreting spatial relations

Spatial relations in specimen locality descriptions are represented as triplets comprising a relatum (the referenced place), a spatial relation, and a locatum (the target location). For example, in the locality description "*18 km west of Ashburton, on Maronan Valetta Road*", *Ashburton* is the relatum, *18 km west of* specifies the spatial relation, and the locatum is the specimen's collection location, which is implicit in the text and represented by the associated coordinates found in the specimen record.

Assuming that the place name to be georeferenced is *Ashburton*, the spatial relation can be inverted, treating the known specimen location as the relatum and the place name as the locatum. Locality descriptions may include additional place names; similar to the example above, "on Maronan Valetta Road" indicates that the specimen was collected on Maronan Valetta Road and does not express a direct spatial relationship between Maronan Valetta Road and Ashburton. Accordingly, our methodology considers only direct spatial relationships between the target place name and the specimen location.

We also acknowledge an asymmetry between relatum and locatum, as described in Talmy's work [25], which highlights differences between the reference object and the object whose location is being described. Reference objects are typically larger, more permanently located, more geometrically complex, and more independent than the objects they help locate. For example, *"the post box is by the post office"* is a more natural expression than *"the post office is by the post box"* because the post office is larger and more permanent, making it a more plausible reference object. However, this asymmetry does not pose a problem for our georeferencing methods, as in our data the locatum is a fixed point observation (the specimen location), and multiple such observations are aggregated to derive spatial constraints.

In our dataset, we observed a wide range of spatial relation terms (Figure 1). The distribution is highly skewed: the proximity relation *near* occurs substantially more frequently than any other term. Cardinal direction relations (*north*, *south*, *east*, and *west*) also appear frequently, whereas the remaining terms occur far less often. In addition, cardinal and

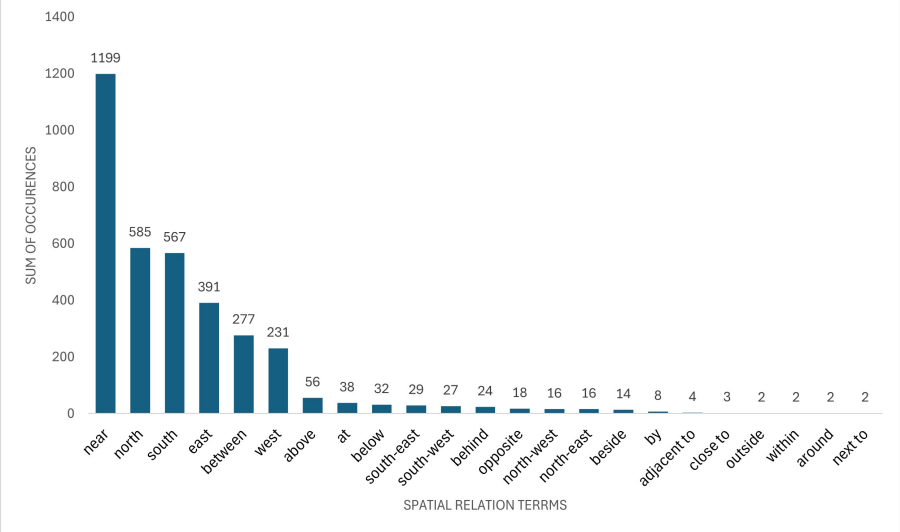


Figure 1 Frequency distribution of spatial relation terms extracted from specimen locality descriptions.

inter-cardinal relations (e.g., south-west, north-east) are sometimes accompanied by an explicit distance, e.g., *"2 km north of"*, which strengthens the spatial constraint beyond direction alone.

Examining the relation terms in our dataset, we attempted to categorise them. However, a key challenge in adopting categorisations from prior studies is that our use case involves inverting the locatum and relatum implied by the locality description; consequently, any categorisation must remain meaningfully interpretable under inversion. While the inverses of cardinal and inter-cardinal relation terms are straightforward, many other relations appear to be symmetrical, such that their interpretation remains unchanged or similar when inverted.

For these non-directional expressions, we therefore treat them as proximity-like constraints even though the perceived degree of proximity could vary. Terms such as *near*, *by*, *close to*, *next to*, *adjacent to*, and *beside* are clearly proximity-related and effectively symmetric (e.g., if X is near Y, then Y is near X). Containment-oriented expressions such as *at*, *within*, and *outside* instead constrain the specimen location relative to a region associated with the referenced feature. However, when the relation is inverted, they do not specify a unique direction or operator for locating the target place name; operationally, they still primarily imply geographic closeness to the relatum. Because specimen records represent the specimen location as a point (latitude/longitude), statements such as "the specimen is within/outside X" (X is a place name) are best interpreted as indicating proximity to X, with the direction left unspecified. Orientation-based terms such as *above*, *below*, *behind*, and *opposite* often depend on a local frame of reference, and thus are not consistently invertible from text alone. Across these cases, modelling non-cardinal relations as proximity-like constraints appears to preserve their role as locality restrictions while avoiding misinterpretation in the inverted setting.

3.3.2 Evaluation dataset design and representation

Pseudo-NGP dataset for evaluation

To evaluate our methods for georeferencing, we constructed a subset from the preprocessed dataset of 13,746 locality descriptions mentioned in Section 3.1. To enable evaluation, we required place names with known coordinates as ground-truth data. Therefore, we selected locality descriptions that contain spatial relations referring to place names recorded in the NZGB Gazetteer. Place names were disambiguated and resolved against the NZGB Gazetteer using the *state/province* information provided in the corresponding biological specimen records. As this study serves as a proof of concept, only place names represented as point geometries were selected in order to simplify the evaluation setting. These gazetteered place names were treated as pseudo-NGPs by withholding their coordinates during inference, allowing us to quantitatively assess the accuracy of the proposed methods. This dataset thus provides a benchmark for systematic evaluation.

Among the resolved place names, we further selected those referenced in at least three locality descriptions, as three constraints constitute the minimum required to define a bounded area. Under this criterion, the final evaluation dataset consists of 2,403 locality descriptions referring to 365 place names. Each locality description is associated with specimen coordinates, and each place name has corresponding resolved ground-truth coordinates.

Input data representation

Both proposed deterministic and probabilistic methods operate over a structured JSON representation derived from the dataset. For each target place name, we extract spatial relation mentions from locality descriptions and group them into a single place-centric entry. Each entry follows the schema *place_name* → *place_name_coordinates* → *relations*:

- **place_name**: identifier for the target place.
- **place_name_coordinates**: ground-truth latitude/longitude of the place_name (retained only for evaluation and not used to form constraints).
- **relations**: a list of spatial relations associated with the *place_name*, extracted from specimen locality descriptions as described in Section 3.1. Each relation entity includes the extracted spatial relation term, an associated distance value when provided, and the coordinates of the specimen record from which the relation was extracted. Relation terms are normalised to a canonical form to reduce surface variation (e.g., collapsing whitespace, standardising hyphenation for multiword expressions such as *next to* → *next-to*, and mapping direction variants such as *NE*, *north east* to a single label). Cardinal and inter-cardinal relations are represented in their inverted form, consistent with our interpretation that the specimen coordinate is the relatum and the target place name is the locatum.

We have used the place name *Kauangaroa*[9] as the running example to illustrate this process. Table 3 lists three locality descriptions that reference *Kauangaroa*, along with the specimen coordinates associated with each description and the extracted (and, where applicable, inverted) spatial relation terms. Listing 1 shows the corresponding JSON entry produced from the data in Table 3. We generated one such JSON entry per place name and combined them into a single JSON file covering all 365 places in our dataset as the input.

[9] *Kauangaroa* is a locality in Wellington, New Zealand.

Listing 1 Example JSON entry used to generate the ALR for the target place name *Kauangaroa*.

```
{
  "place_data": [{
      "place_name": "kauangaroa",
      "place_name_coordinates": {
        "latitude": "-39.928234", "longitude": "175.279616"
        },
      "relations": [
        {
          "relation": "near",
          "distance": "",
          "latitude": "-39.93045", "longitude": "175.2665"
        },
        {
          "relation": "west",
          "distance": "1km",
          "latitude": "-39.92546", "longitude": "175.2909"
        },
        {
          "relation": "east",
          "distance": "",
          "latitude": "-39.91161", "longitude": "175.2635"
        }
    }]
}
```

Building on this representation, we now describe the proposed methods for georeferencing NGPs.

3.3.3 Deterministic georeferencing using spatial relations

In this approach, we framed the georeferencing of NGPs as an inference problem over a *place graph*. The *place graph* provided a compact representation that connects each target place name to all specimen locations whose locality descriptions mention that place name. We then aggregated the resulting spatial constraints to derive an *Approximate Location Region* (ALR) for each target place p. The key steps of this method are as follows:

- **Relatum point set:** For each relation entity in the JSON file generated, we read the specimen coordinates (latitude/longitude) and represented them as relatum points. These points served as the geometric anchors from which constraint regions were generated.

Table 3 Locality descriptions from specimen records that refer to *Kauangaroa* using relative spatial expressions. The latitude and longitude values correspond to the coordinates of the specimen locations described in each locality description.

Place Name	Locality Description	Latitude	Longitude	Extracted Spatial Relation	Inversed Spatial Relation
Kauangaroa	Near Fordell at Kauangaroa	-39.93045	175.2665	near	near (R1)
	1 km east of Kauangaroa, Whangaehu River	-39.92546	175.2909	1 km east	1 km west (R2)
	Whangaehu River catchment west of Kauangaroa	-39.91161	175.2635	west	east (R3)

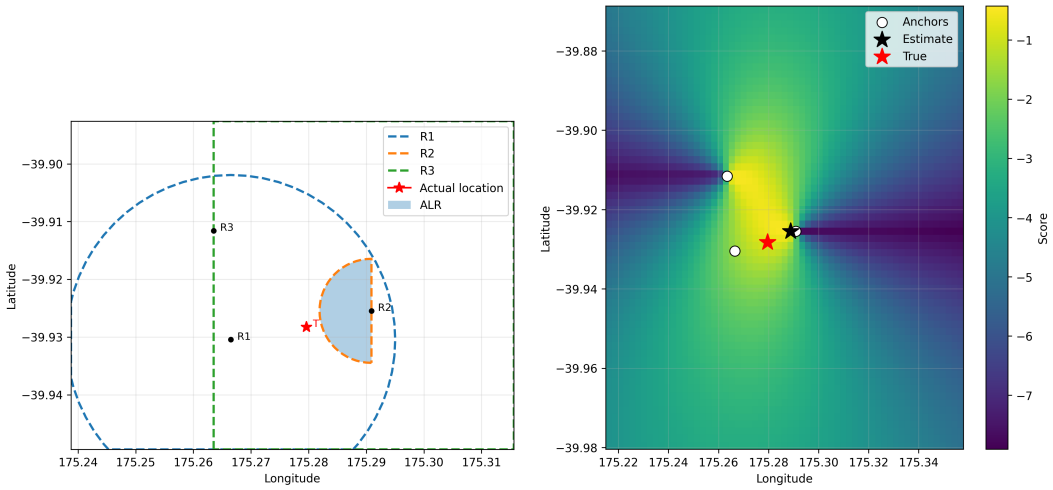
- **Distance parsing:** When a distance value was present, it was parsed and converted to kilometres (from meters or miles). If no distance was provided, it was treated as unspecified and handled according to the relation-specific defaults described below.
- **Spatial context region:** For each target place p, we derived a bounded spatial context to constrain subsequent geometric operations. This context was defined as the bounding box of the relatum point set, expanded by a buffer determined from the median nearest-neighbour distance among the relatum points. This statistic also provided an estimate of the local spread of observations associated with p.
- **Constraint region construction:** Each relation entity was converted into a feasible region within the spatial context:
  - For cardinal and inter-cardinal relations, we generated a directional region by intersecting the spatial context with the corresponding half-plane(s). If a distance was given, the region was additionally intersected with a distance buffer around the relatum point.
  - For all remaining relations, we generated a proximity-like buffer around the relatum point. When an explicit distance was not provided, we set the buffer radius adaptively based on how tightly the specimen (relatum) points for that place name cluster, using their median nearest-neighbour spacing as a proxy for local spread. The resulting default radius was then clipped to global bounds of 0.2 km and 30 km. The lower bound of 0.2 km prevents unrealistically small buffers that would frequently produce empty intersections given coordinate uncertainty and vague locality descriptions, while the upper bound of 30 km prevents weak proximity relations from dominating the spatial context with overly broad regions.
- **ALR computation:** The ALR for p is computed as the intersection of all retained constraint regions.

Figure 2a illustrates the spatial constraint regions generated for *Kauangaroa* from the JSON schema in Listing 1. The figure shows the feasible region associated with each relation (ordered as in the schema) and the resulting ALR obtained as the intersection of these constraint regions. In this formulation, the ALR provides a feasible region rather than a uniquely determined coordinate; accordingly, for evaluation, we used the centroid of the ALR as a point estimate and interpreted the ALR as an uncertainty region. However, an ALR was not obtained for all places: in some cases, the strict intersection became empty due to incompatible or overly restrictive constraints across locality descriptions, particularly when proximity-like relations were imprecise or when extracted relations contain noise. This limitation motivated our second approach, which replaced hard region intersection with a probabilistic formulation to produce a model-derived point estimate.

3.3.4 Probabilistic georeferencing using spatial relations

In this method, we replaced region intersection with a likelihood-based formulation that evaluated candidate coordinates against all extracted spatial relations for a given place name. Each relation contributed a likelihood (soft preference) around its specimen anchor, and the location that maximised the aggregated likelihood was selected as the final coordinate estimate. The key steps of this method are as follows:

- **Relatum point set:** As before, each relation entity provided a specimen coordinate that was treated as the relatum point anchoring the constraint. Distances, when present, were parsed and converted to kilometres.
- **Candidate search space:** For each place name p, we defined a bounded candidate region based on the specimen locations that reference p.



(a) Constrain regions and ALR for *Kauangaroa*. **(b)** Probabilistic score surface for *Kauangaroa*.

Figure 2 Deterministic (a) and probabilistic (b) georeferencing results for *Kauangaroa*. The prediction errors between the actual and estimated locations are 0.71 km and 0.83 km, respectively.

- We computed a robust spatial centre as the median latitude and median longitude of all referencing specimen coordinates.
- We defined a fixed-radius spatial bound (60 km) around this centre to limit the candidate region and reduce unlikely solutions far from the evidence. This radius is approximately four times the estimated proximity uncertainty parameter (s_{near}) described below, beyond which likelihood contributions become negligible. This fixed-radius search assumes that the specimen anchors for a target place form a spatially coherent set. It may be less appropriate for records containing strong outliers or multiple spatial clusters, where an adaptive or cluster-aware search region would be preferable.

- **Likelihood construction:** For each relation, we define a log-likelihood $\ell(x)$ at candidate location x relative to anchor a, where $d(x, a)$ denotes the great-circle (haversine) distance and $\Delta\theta(x, a)$ the angular deviation from the expected bearing; d_0 is the stated distance when provided. Proximity is modelled via distance decay, while directional and distance deviations are modelled using Gaussian (normal) distributions.
  - *Proximity:* $\ell_{\text{prox}}(x) = -d(x, a)/s_{\text{near}}$.
  - *Directional:* $\ell_{\text{dir}}(x) = -\Delta\theta(x, a)^2/(2\sigma_\theta^2)$.
  - *Directional + distance:* $\ell_{\text{dir+dist}}(x) = -\Delta\theta(x, a)^2/(2\sigma_\theta^2) - (d(x, a) - d_0)^2/(2\sigma_d^2)$.

  Here s_{near} (km) controls proximity decay, σ_θ (degrees) directional uncertainty, and σ_d (km) distance uncertainty.
- **Uncertainty parameter estimation:** As described above, the log-likelihood construction depended on three uncertainty parameters: s_{near}, σ_θ, and σ_d. To estimate these values, we used an auxiliary subset of our preprocessed dataset consisting of place names that appeared in only one or two locality descriptions. These place names are listed in the NZGB Gazetteer but were excluded from the evaluation dataset, which requires at least three descriptions per place name. The auxiliary subset contained 1,485 locality descriptions referring to 1,204 resolved place names. Using the ground-truth coordinates of these place names, we computed: (i) anchor-to-ground-truth distances for proximity-like relations, (ii) angular deviations between observed bearings and expected directions

for directional relations, and (iii) differences between observed and stated distances for distance-qualified relations. Because the distance-related distributions included extreme values, we used the 90th percentile of the distance and absolute distance-difference distributions to determine s_{near} and σ_d, thereby representing the majority of cases. For directional relations, we used the median angular deviation to determine σ_θ. The resulting parameter values were $s_{\text{near}} = 14\,\text{km}$, $\sigma_\theta = 47°$, and $\sigma_d = 4\,\text{km}$.

- **Combining evidence:** For a given candidate location, log-likelihoods were summed across all relations to obtain a single score.
- **Coarse-to-fine MAP search:** The bounded region was discretised into an initial coarse grid with 3 km spacing, enabling an efficient first-pass search while preserving sufficient spatial coverage. Each grid point was evaluated by summing relation log-likelihoods, and the best-scoring point was retained. We then iteratively refined the estimate by defining a smaller square window centred on the current best point. The window extent was determined relative to the previous grid spacing, covering a local neighbourhood of candidate cells defined as a square region spanning a fixed multiple of the current grid resolution in each direction. Within this window, progressively finer grid spacings (1.2 km, 0.5 km, and 0.2 km) were applied. At each level, candidates were re-scored, and the maximum was retained. The final latitude/longitude estimate was the highest-scoring point at the finest resolution.

Overall, the method produces a single coordinate for each target place name by selecting the candidate that maximises the aggregated likelihood across all extracted relations. Figure 2b visualises the resulting probabilistic surface for *Kauangaroa*. Warmer colours indicate candidate locations that are more consistent with the extracted spatial relations when evaluated jointly against all specimen anchor points (white markers), and the peak of this surface (black star) corresponds to the model's preferred coordinate estimate.

3.3.5 LLM-based georeferencing

As the third approach, we explored how an LLM performs on this spatial modelling style georeferencing task, using the `gpt-5.1-2025-11-13` model via the OpenAI Batch API[10] with a simple zero-shot prompt. Figure 3a presents the inference prompt; only the locality descriptions and coordinates differ across target places. No system prompt was used, and we did not modify the default model settings.

Because our pseudo-NGP dataset consists of real gazetteered toponyms with coordinates withheld for evaluation, we replaced each target name in the prompt with a placeholder (*Kolombo*) to prevent the model from relying on prior knowledge of the name. The model was explicitly instructed to use only the provided locality descriptions and coordinates as evidence, allowing us to evaluate the ability of a general-purpose LLM to perform spatial modelling from textual spatial information. In a true-NGP setting, however, an LLM may still draw on latent knowledge from historical documents, web pages, maps, or other textual sources seen during training. Such knowledge could improve georeferencing when accurate, but it may also introduce unsupported assumptions. Thus, masking provides a stricter evaluation of inference from specimen-based evidence alone, while true-NGP georeferencing would require further evaluation of how LLMs combine provided evidence with their inherent geographic knowledge.

[10] `https://developers.openai.com/api/docs/guides/batch/`

The LLM was prompted to return a single coordinate estimate for each target place name using the provided locality descriptions and specimen coordinates. To improve transparency, the model was also asked to return its reasoning alongside the predicted coordinates. We found that including this reasoning step also improved the overall accuracy of the LLM-based method, as reported in Table 4. Figure 3b presents the verbatim response from the LLM, including the reasoning behind the inferred coordinates.

(a) Prompt:
Task: Determine the coordinates of "Kolombo".
Constraints:
- Use **ONLY** the locality descriptions and coordinates provided below as evidence.
- Do **NOT** use prior knowledge, assumptions, or any external sources.
- If multiple locations are described, infer Kolombo's position based on their spatial relationships.
- If the evidence is insufficient or contradictory, state that explicitly.

Evidence (multiple locality descriptions):
1) Near Fordell at Kolombo. Lat = -39.93045, Lon = 175.2665
2) 1 km east of Kolombo, Whangaehu River. Lat = -39.92546, Lon = 175.2909
3) Whangaehu River catchment west of Kolombo. Lat = -39.91161, Lon = 175.2635

(b) Response (Verbatim):
We have three points described relative to Kolombo:
1) "Near Fordell at Kolombo": (-39.93045, 175.2665). This is essentially at Kolombo or very close to it.
2) "1 km east of Kolombo, Whangaehu River": (-39.92546, 175.2909). If this point is east of Kolombo, then Kolombo must lie to its west, i.e. at a smaller longitude than 175.2909.
3) "Whangaehu River catchment west of Kolombo": (-39.91161, 175.2635). If this point is west of Kolombo, then Kolombo must lie to its east, i.e. at a larger longitude than 175.2635.
Combining (2) and (3), Kolombo's longitude must lie between 175.2635 and 175.2909. The point explicitly described as "Near Fordell at Kolombo" at 175.2665 falls within this interval, making it a strong candidate.
For latitude, the "at Kolombo" point is -39.93045. The other two sites lie slightly north, consistent with them being nearby but distinct locations.
Thus, the best estimate for Kolombo is approximately -39.930, 175.267 (rounded slightly for representation).

Figure 3 (a) Prompt used to georeference *Kauangaroa* with GPT-5.1. (b) Verbatim response showing the detailed reasoning. The prediction error is 0.8 km.

4 Results and Discussion

To evaluate performance, we report the mean and median error distance (km) between predicted and true coordinates. We also compute A@161 and A@1, measuring the percentage of place names correctly georeferenced within 161 km (100 miles) and 1 km, respectively. A@161 is widely used in toponym resolution and geocoding [8, 12, 23] to capture coarse regional accuracy, while A@1 provides a stricter measure of spatial accuracy.

Table 4 compares the georeferencing performance of the deterministic, probabilistic, and LLM-based methods across these metrics for predicting coordinate estimates for all 365 pseudo-NGP place names in the evaluation dataset. The deterministic model shows the weakest overall performance. It fails to produce estimates for 108 place names, reflecting its reliance on strict rule satisfaction and its inability to handle ambiguous or conflicting spatial constraints. Even when estimates are produced, it records the highest mean error (9.19 km) and the lowest A@1 and A@161 scores.

Both the probabilistic and LLM-based methods successfully generate location estimates for all place names and achieve identical A@161 values, with only two shared outliers exceeding the 161 km threshold. Between these two approaches, the probabilistic model performs best overall, achieving the lowest median error (1.43 km) and the highest A@1 score, indicating more precise fine-grained georeferencing. The two GPT-5.1 prompt variants used for the LLM-based method show broadly comparable performance. Overall, the reasoning prompt performs slightly better in terms of distance-based error, reducing both the mean and the median errors, although the direct prompt achieves a higher A@1 score. The LLM-based variants achieve the lowest mean errors overall, suggesting fewer extreme failures; however, their higher median errors and lower A@1 scores indicate reduced accuracy at finer spatial scales compared to the probabilistic approach.

We examined the two place names for which both the probabilistic and LLM-based methods produced errors exceeding 161 km: Fog Peak and Rockdale. In both cases, the specimen coordinates lie far from the corresponding NZGB gazetteered locations. Fog Peak is an off-track mountain peak, which may contribute to ambiguity in locality descriptions, while the gazetteered location of Rockdale is distant from the specimen coordinates, despite descriptions stating the specimens are "near Rockdale". These discrepancies likely reflect issues in the original data rather than failures of the georeferencing methods.

Table 4 Comparison of the georeferencing performance of the three methods. Mean and median denote error distance in kilometres. ↓ indicates lower is better; ↑ indicates higher is better.

Model	Mean error↓	Median error↓	A@1↑	A@161↑	No estimate↓
Deterministic model	9.19	1.72	20.3%	69.6%	108
Probabilistic model	8.95	**1.43**	**36.0%**	**99.5%**	0
GPT-5.1	6.23	1.85	30.9%	**99.5%**	0
GPT-5.1 with reasoning	**5.89**	1.80	28.1%	**99.5%**	0

4.1 Performance based on the spatial relations

To further investigate the results of the three models, we analysed model performance with respect to the spatial relations present in locality descriptions associated with each place name. Spatial relation terms were categorised into three groups: direction, distance with direction, and proximity, which provided additional insight into the types of relations each model handled more effectively. Table 5 summarises model performance across different combinations of spatial relations. In this table, the LLM-based method refers to the GPT-5.1 reasoning-prompt setting.

The results indicate that although the probabilistic model performs well overall, it is less effective when only directional relations are available, underperforming both the deterministic and LLM-based approaches in this setting. In contrast, when only proximity relations are present, the probabilistic model performs comparatively well. When locality descriptions include all three types of spatial constraints, direction, distance, and proximity, both the probabilistic and LLM-based methods achieve strong performance, with the probabilistic model yielding the highest overall accuracy. Notably, the LLM-based method exhibits relatively consistent performance across different types of relational constraints.

These differences can be partly explained by how each method interprets and integrates different types of spatial relations. Direction-only relations constrain bearing but not distance, leaving the location of the likelihood peak along the directional axis underdetermined. As

Table 5 Comparison of georeferencing performance of the three methods by spatial relation type.

Spatial relations found	Model	Mean error (km)↓	Median error (km)↓
Direction only	Deterministic model	10.02	2.30
Direction only	Probabilistic model	34.58	5.25
Direction only	LLM-based model	**6.46**	**1.41**
Proximity only	Deterministic model	6.87	1.56
Proximity only	Probabilistic model	**7.33**	**1.09**
Proximity only	LLM-based model	7.48	1.32
Direction + Distance	Deterministic model	4.02	2.39
Direction + Distance	Probabilistic model	7.60	**0.74**
Direction + Distance	LLM-based model	**2.59**	2.02
Direction + Proximity	Deterministic model	35.01	2.25
Direction + Proximity	Probabilistic model	8.85	**1.58**
Direction + Proximity	LLM-based model	**7.98**	1.94
Direction + Distance + Proximity	Deterministic model	33.71	1.57
Direction + Distance + Proximity	Probabilistic model	**2.73**	**1.47**
Direction + Distance + Proximity	LLM-based model	2.81	1.82

a result, the probability surface remains diffuse, limiting spatial precision in the absence of additional constraints. Proximity relations, by contrast, align well with probabilistic representations of spatial uncertainty, allowing nearby candidate locations to be weighted without precise orientation. Explicit distance constraints act as strong anchors for all methods, particularly when combined with directional information.

4.2 Key observations on LLM reasoning

We examined the reasoning returned by GPT-5.1 to understand the behaviours that enabled it to georeference effectively across the dataset. Focusing on cases where the LLM outperformed the probabilistic and deterministic methods, two patterns stand out.

First, as shown in Figure 3b, GPT-5.1 is effective at extracting spatial information directly from free text and synthesising it into a coherent interpretation without manual preprocessing. It correctly identifies spatial relations in the text and incorporates associated distance information without relying on an explicit parser. Notably, it independently adopted our approach: inverting the spatial relation so that it is interpreted from the specimen location toward the referenced place name. We also observe that GPT-5.1 accurately performs geographic calculations by combining distance and direction cues, for example, converting stated offsets into rough Δlat/Δlon adjustments and performing basic unit conversions (e.g., miles to kilometres).

Second, the model does not treat each reference equally. Its inference strategy is to find a best-fit solution while ignoring noise, instead of trying to satisfy every relational mention. Rather than weighting all relations equally, the LLM emphasises a subset of mutually consistent (or otherwise more reliable) constraints when deriving an estimate. This behaviour can be observed in the reasoning shown in Figure 3b, in which the LLM primarily uses the "at" coordinates to infer the final coordinates, while the other mentions are used mainly to validate plausibility. This substantially benefits the LLM approach, contributing to the lowest mean error.

4.3 Summary of results

Table 6 provides a summary of the strengths and limitations of each model.

Table 6 Qualitative comparison of georeferencing approaches based on observed behaviour in the experiments.

Method	Strengths	Limitations
Deterministic	Simple and interpretable; Explicitly models an uncertainty region using ALR; Effective when inputs are precise; Exact constraint satisfaction ensures reproducibility.	Requires structured inputs and explicit relation modelling; Sensitive to minor inconsistencies, resulting in the lowest performance.
Probabilistic	Highest overall precision, with the best median error and Acc@1; Performs well when proximity constraints (the most common) are present; Generates estimates for all cases; Reproducible under fixed parameters.	Requires structured inputs and explicit relation modelling; Performs less well with direction-only relations; Affected by inconsistent or imprecise relations.
LLM-based	Operates directly on unprocessed locality descriptions; Produces acceptable estimates for most cases (lowest mean error); Flexible in handling a wide variety of spatial relations (not limited to a pre-defined set).	Reproducibility is not guaranteed; Less precise than the probabilistic model; Costly.

5 Conclusion

This study shows that biological specimen records are a practical source for both identifying and georeferencing non-gazetteered place names within a region. We establish the presence of NGPs in specimen collection records and provide an empirical estimate of their prevalence in the collection. We also compare traditional deterministic and probabilistic georeferencing approaches with a modern LLM-based approach to assess whether LLMs can reliably support a largely spatial modelling–based task. The results indicate that, while the LLM-based method yields broadly acceptable estimates, traditional spatial modelling, particularly probabilistic inference, remains more precise when fine-grained localisation is required.

Limitations. The proposed methods are currently implemented only for NGPs that can be represented as points, limiting applicability to linear or areal features. Performance depends on the availability, spatial distribution, and accuracy of georeferenced specimen anchors, as well as on the diversity of spatial relations available for each place name. The performance of the methods may therefore be biased toward better-constrained place names. In addition, the deterministic and probabilistic methods rely on predefined spatial relation terms and predefined interpretations of those relations. These choices may not fully capture the variability, semantic differences, and context dependence of spatial language. Place name recognition and relation extraction from locality descriptions also remain challenging for deterministic and probabilistic methods due to vague phrasing and inconsistent recording practices, and would benefit from more robust extraction methods.

Future work. Future extensions will support non-point geometries (lines and polygons) and richer representations of uncertainty. We also aim to improve spatial relation parsing to capture both explicit and implicit relations. The LLM-based method could also be further improved through more systematic prompt design, including in-context learning with few-shot examples, and exploration of uncertainty estimation from model outputs. We further aim to develop a hybrid pipeline that combines the strengths of LLMs with probabilistic inference to georeference NGPs. Finally, we will expand the analysis to additional regions, particularly those with limited historical coverage in existing gazetteers, alongside improved methods for NGP recognition.

References

1 Elise Acheson, Stef De Sabbata, and Ross S. Purves. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017. doi:10.1016/j.compenvurbsys.2017.03.007.

2 Edwin Aldana-Bobadilla, Alejandro Molina-Villegas, Ivan López-Arévalo, Shanel Reyes-Palacios, Victor Muñiz-Sanchez, and Jean Arreola-Trapala. Adaptive geoparsing method for toponym recognition and resolution in unstructured text. *Remote Sensing*, 12(18):3041, 2020. doi:10.3390/rs12183041.

3 Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander. Geoparsing historical and contemporary literary text set in the city of edinburgh. *Language Resources and Evaluation*, 53(4):651–675, 2019. doi:10.1007/s10579-019-09443-x.

4 Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander, and Richard Tobin. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, pages 3936–3944, 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/129.html.

5 Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS Int. J. Geo Inf.*, 11(1):28, 2022. doi:10.3390/ijgi11010028.

6 Arthur Chapman and John Wieczorek. Georeferencing best practices, 2020. doi:10.15468/DOC-GG7H-S853.

7 Hao Chen, Maria Vasardani, and Stephan Winter. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 17(1):31–62, 2018. doi:10.5311/JOSIS.2018.17.417.

8 Grant DeLozier, Jason Baldridge, and Loretta London. Gazetteer-independent toponym resolution using geographic word profiles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2382–2388, February 2015. doi:10.1609/aaai.v29i1.9531.

9 Department of Conservation. Banks peninsula conservation walks, 2011. URL: https://www.doc.govt.nz/globalassets/documents/parks-and-recreation/tracks-and-walks/canterbury/banks-peninsula-conservation-walks.pdf.

10 Aneesha Fernando, Surangika Ranathunga, Kristin Stock, Raj Prasanna, and Christopher B. Jones. Georeferencing complex relative locality descriptions with large language models. *International Journal of Geographical Information Science*, 0(0):1–33, 2026. doi:10.1080/13658816.2026.2613291.

11 Jacques Fize, Ludovic Moncla, and Bruno Martins. Deep learning for toponym resolution: Geocoding based on pairs of toponyms. *ISPRS Int. J. Geo Inf.*, 10(12):818, 2021. doi:10.3390/ijgi10120818.

12 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, 2018. doi:10.1007/s10579-017-9385-8.

13 Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014. doi:10.1613/jair.4200.

14 Linda L. Hill. *Georeferencing - The Geographic Associations of Information*. Digital libraries and electronic publishing. Mit Press, 2009.

15 Xuke Hu, Jens Kersten, Friederike Klan, and Sheikh Mastura Farzana. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, 40(3):670–697, 2026. doi:10.1080/13658816.2024.2405182.

16 Ehsan Kamalloo and Davood Rafiei. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1287–1296. ACM, 2018. doi:10.1145/3178876.3186027.

17 Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B. Jones. Georeferencing wikipedia documents using data from social media sources. *ACM Trans. Inf. Syst.*, 32(3):12:1–12:32, July 2014. doi:10.1145/2629685.

18 Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. GeoLM: Empowering language models for geospatially grounded language understanding. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.317.

19 Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. In *Proceedings of the 25th AGILE Conference on Geographic Information Science, Vilnius, Lithuania, July 18-22, 2022*, volume 3, page 9. Copernicus Publications Göttingen, Germany, 2022. doi:10.5194/agile-giss-3-9-2022.

20 Kai Ma, Yongjian Tan, Zhong Xie, Qinjun Qiu, and Siqiong Chen. Chinese toponym recognition with variant neural structures from social media messages based on BERT methods. *Journal of Geographical Systems*, 24(2):143–169, 2022. doi:10.1007/s10109-022-00375-9.

21 Arnald Marcer, Elspeth Haston, Quentin Groom, Arturo H Ariño, Arthur D Chapman, Torkild Bakken, Paul Braun, Mathias Dillen, Marcus Ernst, Agustí Escobar, et al. Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. *Diversity and Distributions*, 27(3):564–567, 2021.

22 Jamie Scott, Kristin Stock, Fraser Morgan, Brandon Whitehead, and David Medyckyj-Scott. Automated georeferencing of antarctic species. In *11th International Conference on Geographic Information Science, GIScience 2021, Poznań, Poland (Virtual Conference), September 27-30, 2021 - Part II*, volume 208 of *LIPIcs*, pages 13:1–13:16. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.GIScience.2021.II.13.

23 Praval Sharma, Ashok Samal, Leen-Kiat Soh, and Deepti Joshi. A spatially-aware data-driven approach to automatically geocoding non-gazetteer place names. *ACM Transactions on Spatial Algorithms and Systems*, 10(1):1–34, 2024. doi:10.1145/3627987.

24 Philip David Smart, Christopher B. Jones, and Florian A. Twaroch. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Geographic Information Science, 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings*, volume 6292 of *Lecture Notes in Computer Science*, pages 234–248. Springer, Springer, 2010. doi:10.1007/978-3-642-15300-6_17.

25 Leonard Talmy. Toward a cognitive semantics, vol. 1 (concept structuring systems), 2000.

26 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian van der Meij. Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4):563–581, 2015. doi:10.1111/tgis.12110.

27 Jimin Wang and Yingjie Hu. Are we there yet?: evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, GeoHumanities@SIGSPATIAL 2019, Chicago, IL, USA, November 5, 2019*, pages 2:1–2:6. Association for Computing Machinery, 2019. doi:10.1145/3356991.3365470.

28 John Wieczorek, Qinghua Guo, and Robert J. Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International journal of geographical information science*, 18(8):745–767, 2004. doi:10.1080/13658810412331280211.

29 Yibo Yan and Joey Lee. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4163–4167. ACM, 2024. doi:10.1145/3627673.3679934.

30 Madiha Yousaf and Diedrich Wolter. A reasoning model for geo-referencing named and unnamed spatial entities in natural language place descriptions. *Spatial Cognition & Computation*, 22(3-4):328–366, 2022. doi:10.1080/13875868.2021.2002872.

31 Bing Zhou, Lei Zou, Yingjie Hu, Yi Qiang, and Daniel W. Goldberg. Topobert: a plug and play toponym recognition module harnessing fine-tuned BERT. *International Journal of Digital Earth*, 16(1):3045–3064, 2023. doi:10.1080/17538947.2023.2239794.

Extending ISO-Space to Annotate Spatial Information in Route Descriptions

Samira Goudarzi Ali Akbari[1]
The University of Melbourne, Parkville, Australia

Stephan Winter
The University of Melbourne, Parkville, Australia

Martin Tomko
The University of Melbourne, Parkville, Australia

Abstract

A robust spatial annotation scheme that explicitly represents spatial elements and the links between them in English natural (human) language route descriptions is required to support computational interpretation. Yet, existing spatial annotation schemes for English do not capture all the spatial elements and links that occur in route descriptions, including perceptual expressions and reorientation actions. Using Formal Concept Analysis (FCA), the elements and links are derived by modeling the conceptual space of route descriptions. These elements and links are then compared against ISO-Space, an annotation standard for encoding spatial entities, motion events, and spatial relations [28], selected as the reference framework due to its relatively comprehensive coverage among existing schemes. Based on the analysis, targeted scheme refinements and extensions interoperable with ISO-Space are introduced. The resulting scheme is evaluated through two annotation rounds on 48 route descriptions drawn from four datasets. The results show improved consistency and increased annotation coverage through perceptual and reorientation links, which account for 23% of all annotated links, without compromising annotation reliability.

2012 ACM Subject Classification Computing methodologies → Language resources; Computing methodologies → Ontology engineering

Keywords and phrases spatial language, spatial information annotation scheme, route description

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.12

1 Introduction

People use verbal descriptions to communicate spatial information, thus externalizing their mental representations of their spatial environment. A route description is defined as "*a verbal instruction to follow a particular route through an environment*" [37, p. 4]. In real-world scenarios, such as emergency calls, route descriptions may be the only information available to guide someone to a specific location. The computational interpretation of such instructions supports applications including the conversion of verbal descriptions into sketch-map representations [34, 6] and geocoding [21]. These applications rely on the automatic extraction of spatial information from route descriptions. Such extraction typically involves machine learning models [30, 15, 27], which in turn require an annotated corpus for training.

To train a model for the automatic extraction of spatial information in route descriptions, we first need an annotation scheme that specifies which spatial elements and links to annotate. Consider the following example of a human-generated route description: *Walk straight along Park Street, and turn left when you see the library on your right. Follow the path past the stone bridge, then turn left at the fountain.* According to Denis [8], landmarks (e.g., *library*)

[1] Corresponding author

© Samira Goudarzi Ali Akbari, Stephan Winter, and Martin Tomko;
licensed under Creative Commons License CC-BY 4.0

and action descriptions constitute two critical elements of route descriptions. Denis further identifies two main classes of actions: changing orientation (e.g., *turn left*) and proceeding (e.g., *walk straight along*). In addition to the two main actions, Denis identifies another class of action, positioning and inception (e.g., *turn left when you see the library on your right*), in which perception verbs (e.g., *see*) participate [2]. Perceptual expressions allow the moving individual to check their current location and orientation, often based on descriptions of the visual scene, whereas changing orientation actions introduce decision points at which orientation needs to be changed. Perceptual expressions and changing orientation actions must be captured in an annotated dataset of route descriptions.

Generating annotated datasets requires robust annotation schemes that define which spatial elements and links can be consistently recognized across corpora. Previous research has proposed annotation schemes aimed at identifying spatial information in language [20, 16, 28] based on broader research in spatial semantics [31, 13]. None of these schemes fully captures the spatial elements that need to be annotated in route descriptions. For example, SpatialML [20] focuses primarily on the annotation of place elements but does not address the proceeding actions. Spatial Role Labeling [16] captures roles, but does not represent the full semantics of spatial relations expressed in spatial discourse. Pustejovsky et al. [28] go further by providing the ISO-Space annotation scheme to encode spatial entities, motion events, and spatial relations, aiming to ameliorate the limitations of earlier standards. However, critical elements frequently occurring in route descriptions, including perceptual expressions and reorientation actions, are still not explicitly represented through dedicated tags in the current ISO-Space specification. These gaps highlight the need for an annotation scheme that captures the spatial elements and links present in verbal route descriptions.

In this paper, three research questions (RQs) are addressed: **RQ1**: What conceptual elements and links are required to achieve a systematic representation of spatial information in walking route descriptions beyond what ISO-Space currently provides? **RQ2**: How can the unrepresented elements and links be formally defined and integrated into the ISO-Space framework without compromising its ontological coherence and interoperability? **RQ3**: How does the extended annotation scheme improve coverage and consistency compared with the original ISO-Space specification when applied to English route corpora?

To address RQ1, a systematic methodology based on Formal Concept Analysis (FCA) is used to model the conceptual space of spatial discourse and identify conceptual gaps in ISO-Space through a theory-driven lattice structure. This step demonstrates not only the conceptual gaps in ISO-Space scheme but also the need for refinements. For RQ2, each uncovered and refined spatial concept is formally defined and integrated into the ISO-Space framework using the CASCADES stages [7], ensuring ontological coherence and interoperability. Finally, RQ3 evaluates the extended scheme through two annotation rounds applied to three walking route corpora and one hiking route corpus: the first annotation round using the original ISO-Space standard and the second annotation round using the systematically extended standard, to assess improvements in coverage and consistency. These corpora represent naturalistic route descriptions, providing a suitable testbed to evaluate annotation coverage, and consistency. All datasets are in English, which limits the scope of this study to avoid cross-linguistic variation in spatial relations and grammatical patterns. The contributions of this study are therefore:

- Systematic identification of spatial elements and links in route descriptions that are not represented in the ISO-Space annotation standard, including reorientation actions and perception expressions;
- Systematic identification of spatial elements and links in route descriptions that require refinement in the ISO-Space annotation standard to improve coverage and consistency;

- Design of targeted extensions and refinements that explicitly encode missing elements and links, together with a formal XML specification that integrates these into a scheme compatible with the ISO-Space framework; and
- Application of the extended scheme to three corpora of walking route descriptions and one hiking corpus, demonstrating increased annotation coverage and consistency.

Section 2 reviews the conceptual foundations of route descriptions and existing spatial annotation schemes. Section 3 describes the datasets used to construct the annotation corpus. Section 4 applies Formal Concept Analysis to address the first research question. Section 5 details the integration of the resulting refinements and extensions into the ISO-Space XML representation, addressing the second research question. Section 6 describes the evaluation procedure and reports the results addressing the third research question. Section 7 concludes the paper and discusses implications, limitations, and future work.

2 Related work

This section provides an overview of the relevant literature. Section 2.1 examines the conceptual components of route descriptions and Section 2.2 reviews existing spatial annotation schemes.

2.1 Route description components

Route directions constitute a particular form of spatial discourse whose purpose is to guide an individual through an unfamiliar environment to a specified destination [9]. Other studies have identified the components involved in route descriptions and have proposed definitions for these components:

Segment. According to Tversky et al. [33], a segment is a conceptual unit that divides route depictions and descriptions into start point, reorientation, path/progression, and end point, which contains sufficient information to allow movement between points. Jackendoff's path constituents explicitly include start and end points of motion and decompose motion as a moving entity that traverses a path over a specific time period [12]. Allen [2] identifies pathways as environmental features that denote actual or potential channels of movement.

Decision point. Locations that offer alternative paths are typically exemplified by intersections [2]. In cities, a node is a strategic point or focus that a moving individual can enter and from which movement within the city is organized [19]. Nodes are junctions, which function as navigation decision points, or concentrations, which derive importance from the condensation of a specific use or physical character.

Landmark. Landmarks are point-references external to the moving entity, which does not enter them [19]. Based on Lynch, Allen [2] defines landmarks as environmental features that serve as point-references and intermediate goals along a path of movement from the origin to the destination. Denis [9] further characterizes landmarks as three-dimensional physical objects, either natural (e.g., *church*) or objects intended to convey directions (e.g., *signposts*), as well as two-dimensional entities (e.g., *streets*).

Action description. Denis [8] distinguishes two main action classes: proceeding (e.g., *walk straight ahead*) and changing orientation (e.g., *turn left*). Additional actions include positioning and inspection, which allow moving entities to check orientation based on the visual scene (e.g., *when you arrive there you will see the library on your right*) [8] and involve perception verbs (e.g., *see*) [2]. the proceeding and changing orientation are expressed by movement verbs, defined by Allen [2] as statements prescribing where the moving entity should go and semantically reducible to *go* or *turn*.

Unlike Allen, Winter et al. [37] distinguish verbs denoting proceeding actions from those expressing changes in orientation, the latter of which are not integrated into motion verb classes proposed in other literature. Muller [22] introduces six motion verb classes, including *leave, hit, reach, external, internal, and cross* within a qualitative theory of motion based on spatio-temporal primitives. Asher et al. [3] propose ten classes of motion verbs in French, including *s'approcher* (approach), *arriver* (arrive), *entrer* (enter), *se poser* (alight), *s'éloigner* (move away), *partir* (leave), *sortir* (exit), *décoller* (take off), *passer (par)* (pass through), and *dévier* (deviate). Pustejovsky et al. [29] map the motion classes proposed by Muller, Asher et al., and Baker et al. [5], resulting in ten unified motion verb classes: *move, move-external, move-internal, leave, reach, detach, hit, follow, deviate, and stay.*

Descriptive component. Spatial relations between environmental features or between moving entities and environmental features can be provided in route descriptions through direction designations [2] or topological relations [8]. This form of spatial description is conveyed using state-of-being verbs that are reducible to *is* [2]. Direction designations specify spatial relations using abstract, object-based, or body-based frames of reference [2].

Distance designation. Spaces that separate points of reference are expressed through standard units, conventional nonstandard units, temporal units, or vague judgments [2], and are expressed qualitatively rather than quantitatively [32].

This review of route description components establishes a theoretically grounded set of conceptual building blocks from prior research, which provides the basis to specify linguistic annotations units for evaluating the representational coverage of ISO-Space.

2.2 Spatial information annotation schemes

SpatialML is a framework developed to support the annotation of spatial information [20], mainly aimed at marking spatial locations for integration with external resources such as domain-specific databases and gazetteers. SpatialML does not represent motion events because it does not annotate predicates.

Kordjamshidi et al. [16] introduce Spatial Role Labeling (SpRL), a spatial annotation scheme grounded in Holistic Spatial Semantics (HSS) [38] that identifies spatial arguments and relations. SpRL models static and dynamic relations using three roles, including trajector, landmark, and spatial indicator, formally represented as semantic triples [16].

TimeML is created by Pustejovsky et al. to represent temporal information in text [26]. Pustejovsky et al. [29] discuss how the structure of spatial events motivates spatio-temporal markup, porposing SpatialML. ISO-Space, designed to capture spatio-temporal and spatial information in natural language, builds on the principles of TimeML and SpatialML [28]. Although ISO-Space aligns with SpRL, ISO-Space incorporates more detailed and fine-grained spatial and linguistic elements.

The detailed semantics of ISO-Space complicate data preparation for machine learning [15]. Yet, Pustejovsky et al. [28] argue for this detail. Simple schemes, such as SpRL may fail to capture the intended interpretation of spatial expressions. For example, assigning the

topological relation Tangential Proper Part (TPP) to *The girl is in the room* does not clearly convey the intended meaning and may suggest an unintended interpretation, such as the girl being attached to the side of the room rather than standing on the floor inside it.

The latest version of ISO-Space (ISO 24617-7) provides an XML-based spatial information annotation scheme to represent spatial elements and links. The scheme defines elements for spatial entities, events, spatial relations, and dimensional information, the latter encoded using *<measure>*. Spatial entities comprise three subtypes: *<place>*, *<path>*, and *<eventPath>*. While *<place>* and *<path>* represent locational entities, *<eventPath>* denotes a trajectory created by a motion event. In addition to these locational or trajectorial entities, ISO-Space also allows for non-locational *<entity>* that function as a spatial entity or participates in motion events. Events are encoded using *<event>* and include both motion and non-motion events, with *<motion>* representing events that involve a change of location of an *<event>*. Spatial relations are encoded using *<sRelation>*, which includes attributes of *topological*, *directional*, *pathDefining*, and *goalDefining*. the topological and directional relations between elements are expressed using the link types *<qsLink>* and *<oLink>*, where *<oLink>* additionally incorporates an explicit frame of reference and reference point. Motion relations are encoded using *<moveLink>*, which relates a moving entity to the *<eventPath>* it traverses, while *<mLink>* associates a *<measure>* with the entity being measured.

ISO-Space (ISO 24617-7) is the only ISO standard for the annotation of spatial information. Both the standard and its accompanying corpus (SpaceEval) have been adopted in research on Natural Language Processing (NLP) to extract spatial relations [23, 30, 35, 24]. Despite this progress, it has not gained widespread adoption due to its complexity and the lack of annotated datasets and automated tagging tools [11]. This paper shows that this complexity does not hinder implementation and demonstrates that the proposed extensions and refinements improve coverage and consistency in annotating elements and links in route descriptions.

3 Data

This section details the data foundation and is divided into two subsections. Section 3.1 outlines the selection requirements to identify suitable datasets and introduces the selected datasets, describing their content and characteristics. Section 3.2 describes the procedure for compiling a representative corpus of 48 route descriptions.

3.1 Dataset selection and description

This research uses datasets of authentic route descriptions produced by different users, covering activity types (e.g., *walking, hiking*), geographic contexts (e.g., *urban, natural and rural*), and environmental settings (e.g., *indoor, outdoor*) to ensure generalization. Four datasets that meet these criteria are identified: *Walkingmaps*, *Slowways*, *Soleway*, and *Summitpost*. The first three describe walking activities, while *Summitpost* includes hiking and mountaineering activities. *Walkingmaps*, *Slowways*, and *Summitpost* represent outdoor route descriptions in rural, urban, and natural areas, whereas *Soleway* contains indoor ones.

- *Walkingmaps* (`https://walkingmaps.com.au/`) is a publicly accessible platform operated by Victoria Walks Inc., a non-profit organization that promotes walkable communities in Australia. It enables users to discover and share leisure walks. Each walk includes a trajectory, points of interest with verbal descriptions and optional images, and a textual description that provides navigational guidance and experience, along with metadata such as distance, estimated duration, and difficulty. Web scraping was used to extract 2,631 walks from the *Walkingmaps* website. The dataset was collected in July 2025.

Table 1 Size of the sampling frame across all datasets.

	Walkingmaps	Slowways	Soleway	Summitpost	Total
Count	2495	16229	1156	1795	21675

- The *Slowways* project (`https://beta.slowways.org/`) is a citizen-created national walking network connecting Britain's towns, cities, and national landscapes. Each route page includes an *Overview*, *Surveys*, *Geography*, and *Reviews*, with the option to download route geometry. The *Reviews* section provides textual descriptions with navigational information, while the *Overview* lists start and end point names, photos, and details. The *Surveys* section reports facilities, accessibility, obstacles, challenges, and key measurements, and the *Geography* section provides quantitative details such as length and elevation. Access to the dataset was granted in August 2025, comprising 16,450 routes.
- *Soleway* (`https://soleway.ugent.be/`) is a crowd-sourced indoor navigation service that offers wayfinding capabilities through textual route descriptions, avoiding expensive technologies, and promoting ease of use. Each route description includes metadata such as author name, author id, and created date. The dataset used in this study corresponds to the state of the *Soleway* platform in 2018 and contains 1,164 route descriptions.
- *Summitpost* (`https://www.summitpost.org/`) is a collaborative platform based in the united states that focuses on outdoor activities such as climbing. The search parameters for route entries include location (e.g., *continent*), route type (e.g., *hiking*), and rock difficulty. Route entries comprise sections such as *Overview*, *Getting There*, *Route Description*, *Essential Gear*, *Commentary*, *When to Climb*, and *Images*. Each route includes metadata such as coordinates, season of use, estimated duration, and difficulty ratings. In October 2025, 1,795 hiking route records met the search constraints of continent (North and South America), country (United States), and route type (hiking); these constraints were applied to ensure diversity of geographic contexts and activity types.

3.2 Corpus construction for annotation

The sampling frame consists of all available route descriptions from the four datasets. After excluding records that do not contain route descriptions, the final sampling frame comprises 21,675 descriptions, as summarized in Table 1. Systematic stratified sampling is used to capture the variability between heterogeneous strata of route descriptions from different datasets. The sampling frame is partitioned into mutually exclusive *strata*, and samples are drawn from each stratum so that all subgroups are represented. The stratified sampling procedure by which the final corpus of 48 route descriptions is derived from the sampling frame is described below.

- Each description is categorized by two criteria: geographic setting (urban, natural/rural, or indoor) and activity type (walking or hiking). The sampling frame is divided into strata defined by the cross-product of these criteria, yielding six strata. The indoor–hiking stratum is excluded as conceptually meaningless, while the urban–hiking stratum is removed due to sparse representation. The remaining strata are internally homogeneous with respect to geographic setting and activity type and together partition the sampling pool, ensuring valid combinations. The number of route descriptions in each stratum is summarized in Table 2.
- A quota-based sampling strategy is adopted, with a fixed quota of 12 route descriptions selected from each stratum, resulting in a corpus of 48 route descriptions (711 sentences). This corpus size supports reliable evaluation metrics for the training of computational

Table 2 Number of route descriptions in each stratum.

Stratum	Walkingmaps	Slowways	Soleway	Summitpost	Total
urban–walking	1656	0	0	0	1656
natural/rural–walking	839	16229	0	0	17068
indoor–walking	0	0	1156	0	1156
natural/rural–hiking	0	0	0	1795	1795

Table 3 Summary statistics for words and sentences in the 12 samples of the four strata.

	Words				Sentences			
Stratum	Count	Min	Mean	Max	Count	Min	Mean	Max
urban–walking	3396	114	283	486	260	12	22	39
natural/rural–walking	3032	117	253	472	209	10	17	42
indoor–walking	546	26	46	91	68	2	6	13
natural/rural–hiking	2432	70	203	361	174	4	15	26

models [30] and is consistent with the corpus sizes used to develop gold standard datasets [27, 1]. In proportionate allocation, the sample size per stratum is aligned with its relative size to preserve the population distribution and minimize bias. However, since more than 78.7% of the population falls within a single stratum, this approach is not suitable to achieve an interpretable sample. Optimum allocation requires prior knowledge of the variance per stratum, which, for textual data, depends on previously annotated elements that are unavailable. Therefore, minimum quotas are selected to ensure balanced representation across strata and avoid over-representing the dominant stratum.

- Within each stratum, random sampling is applied to select 12 route descriptions objectively and without bias. No contributor appears more than once in the final corpus; any violating description is replaced by another random one drawn from the same stratum. Table 3 presents summary statistics for the 12 route descriptions selected from each of the four strata. Count reports the total number of words or sentences per stratum, while Min, Mean, and Max summarize the per-description word and sentence counts. Indoor walking descriptions tend to be shorter because they focus on wayfinding within confined spaces, while outdoor walking, often undertaken as a leisure activity, typically takes hours and, consequently, involves longer route descriptions.

4 Scheme development via Formal Concept Analysis

To address RQ1, a systematic approach is required to explore the conceptual space of route descriptions. Formal Concept Analysis (FCA) is a mathematical theory of concept formation derived from lattice and ordered set theories that reveals relationships between objects and attributes [36]. FCA operates on a formal context, defined as a set of incidence relations between a set of objects and a set of attributes, indicating which objects possess which attributes. Its main construct is the *formal concept*, defined by its *extent* (the set of objects) and *intent* (the set of shared attributes). Formal concepts are organized through a generalization–specialization ordering that forms a concept lattice [36, 10]. Concept lattices can be represented as Hasse diagrams, which provide a structured visualization of conceptual spaces and support the systematic identification of conceptual gaps [10].

The process of applying FCA to answer RQ1 begins with a conceptual analysis in Section 4.1, which specifies the semantic requirements of the route descriptions. Building on this foundation, Section 4.2 describes the construction of formal contexts for both *elements*

and *links*, followed by Section 4.3, which details the generation of the corresponding concept lattices. The resulting structures are then used in Section 4.4 to propose refinements and in Section 4.5 to extend newly introduced elements and links that address the conceptual gaps identified through FCA.

4.1 Conceptual analysis

The linguistic units in the annotation scheme are either *elements*, representing semantic arguments in natural (human) language route descriptions, or *links*, expressing relations between those elements. Conceptual analysis serves as the primary analytical stage, specifying the linguistic units required to capture route description components. In the following, the linguistic units, including elements and links, are defined.

- **Elements**:
  place: A word(s) that refers to an entity resulting from a shared identification of a location. This definition is based on the derivation that reflects an ontological commitment introduced by Purves et al. [25]. This linguistic unit is used to annotate references to landmarks and decision points based on the notion of Allen [2].
  path: A word(s) that denotes a linear entity that functions as an actual or potential traversable channel. This linguistic unit is used to annotate references to paths following Allen [2].
  moving entity: A word(s) that denotes the entity undergoing a change of location.
  spatial relation: A word(s) that indicates how two entities are located relative to each other. In syntactic terms, spatial relations are most often represented using prepositions, but not always. This linguistic unit is used to annotate references to direction designations [2] or topological relations [8].
  motion: A verb that indicates an event in which the moving entity moves from a source, along a path, and towards a goal. This linguistic unit is used to annotate references to motion verbs [29] in proceeding actions [8].
  perception: A verb that indicates an event in which the moving entity observes or notices an entity, using sensory information to confirm location or orientation. This linguistic unit is used to annotate references to perception verbs [37] in inspection and positioning actions [8].
  reorientation: A verb that indicates an event in which the moving entity changes facing direction without displacement or with displacement, typically at decision points. This linguistic unit is used to annotate references to reorientation verbs [37] in changing orientation actions [8].
  event: A word(s) that indicates a broad notion of an event that includes all kinds of actions, states, processes, activities, and accomplishments [4]. This linguistic unit is used to annotate references to events, except perception, reorientation, and motion events.
  measure: A word(s) that indicates the degree of separation between two reference points qualitatively and quantitatively and applies to both spatial and temporal dimensions. This linguistic unit is used to annotate references to distance designations [2].
- **Roles**:
  figure: An entity whose location, orientation, or motion is described.
  ground: An entity relative to which the location, orientation, or motion of the figure is specified.
- **Links**:
  topological link: A link that associates a *figure* with a *ground* through a topological spatial relation. This linguistic unit is used to relate elements in descriptive components [2].

directional link: A link that associates a *figure* with a *ground* through a directional spatial relation within a frame of reference (absolute, intrinsic, relative). This linguistic unit is used to relate elements in descriptive components [2].
movement link: A link that associates a motion event with its *figure*, *ground*, and involved elements of *source* (path or place that is the origin of the movement), *path* (path in which movement occurs), *mids* (path or place that is intermediate goals of movement), *spatial relation*, and *measure*. This linguistic unit is used to relate the elements in proceeding actions [8].
reorient link: A link that associates a reorientation event with its *figure*, *ground*, and involved elements of *source* (the place or path at which the reorientation occurs), and *spatial relation*, indicating the new facing direction. This linguistic unit is used to relate the elements in changing orientation actions [8].
perceive link: A link that associates a perception event with its *figure* and *ground*, recording sensory confirmation of a *ground*. This linguistic unit is used to relate elements in inspection and positioning actions [8].

The defined units guide the determination of the sets of objects and attributes required to construct the formal contexts, with elements and links treated separately.

4.2 Formal context construction

The construction of formal contexts begins by determining the sets of objects and attributes required to represent the units defined in Section 4.1 separately for elements and links. The objects constitute the rows, and the attributes constitute the columns of the context matrices. Each cell is assigned the value 1 when an object exhibits a given attribute and 0 otherwise.

All meaningful combinations of attributes are systematically generated to support comparison across objects and to identify three outcomes: combinations missing from ISO-Space, combinations that correspond to existing ISO-Space elements and links, and combinations for which an ISO-Space counterpart exists but requires refinement. Semantically meaningless combinations are excluded. For missing combinations, a new object name is introduced; for corresponding combinations, the ISO-Space name is adopted; and for incomplete combinations, the ISO-Space name is retained, and the object is treated as a refinement.

To ensure a consistent and readable presentation, a uniform notation is adopted for all objects and attributes in the FCA. The names of objects, whether *elements* or *links*, are noted using angle brackets (e.g., *<entity>*, *<qsLink>*) and attributes are prefixed with @ (e.g., *@event*, *@spatialrelation*), following the convention used in the ISO standard. To distinguish the contributions of this work from the original ISO-Space specification, all newly introduced or modified attributes and objects are marked with the *rs:* prefix (***R****oute* ***S****emantics*), following standard XML conventions. For example, while *<qslink>* remains an ISO-Space link, *<rs:reorientLink>* represents a new link introduced in this work. In the following, the construction of the formal contexts for both elements and links is described. The modified ISO-Space scheme, incorporating both the proposed extensions and refinements, is referred to as *ISO-Space+RP*, where R denotes reorientation and P denotes perception.

- **Elements:** A set of attributes is selected to conceptually characterize the semantics of objects. Attributes already present in ISO-Space, such as *@spatial entity*, *@locational-one*, *@locational-series*, *@non-locational*, *@traversal*, *@event*, *@motion event*, *@motionClass*, *@non-motion event*, *@spatial relation*, *@topological*, *@directional*, *@pathDefining*, and *@goalDefining*, are reused whenever they correspond to a required semantic distinction. Two further distinctions of events required lack ISO-Space counterparts. Therefore, we

Table 4 Formal context for elements.

	@spatial entity	*@locational-one*	*@locational-series*	*@non locational*	*@traversal*	*@event*	*@motion event*	*@motionClass*	*@non motion event*	*@rs:reorientation event*	*@rs:perception event*	*@spatial relation*	*@topological*	*@directional*	*@pathDefining*	*@goalDefinig*
<entity>	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
<rs:place>	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<rs:path>	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
<eventPath>	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
<motion>	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
<event>	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
<sRelation>	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
<rs:reorientation>	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
<rs:perception>	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0

introduce two additional attributes *@rs:reorientation event* and *@rs:perception event* to ensure that all meaningful attribute combinations required to represent elements in the route descriptions can be captured systematically.

The objects *<entity>*, *<rs:place>*, *<rs:path>*, *<eventPath>*, *<motion>*, *<event>*, and *<sRelation>* of ISO-Space are semantically aligned or require modification to align with the required elements defined in Section 4.1. The meaningful attribute combinations and conceptual analysis show that route descriptions introduce additional events, such as *<rs:reorientation>* (e.g., *turn* in *turn left at the fountain*) and *<rs:perception>* (e.g., *see* in *you see the library on your right*), for which ISO-Space provides no specific elements. Table 4 shows the formal context of the elements.

- **Links:** For attributes, whenever a required distinction already exists in ISO-Space, its attribute name is reused; these include *@event*, *@spatialRelation*, *@figure*, *@ground*, *@trelType*, *@drelType*, *@frameType*, *@mrelType*, *@motion*, *@sRelation*, *@merelType*, *@measure*, and *@bound*. When conceptual analysis reveals distinctions that lack ISO-Space counterparts, a new attribute name is introduced. Consequently, two additional attributes, *@rs:reorientation* and *@rs:perception*, are added to the attribute set to enable the construction of *<rs:reorientLink>* and *<rs:perceiveLink>*.

  For objects, combinations of attributes that fully correspond to existing ISO-Space links adopt the ISO-Space name, as in *<qsLink>*, *<oLink>*, and *<mLink>*. Combinations that partially correspond retain the ISO-Space name but are treated as refinements, as illustrated by *<rs:moveLink>*. Combinations without an ISO-Space counterpart require new links, motivating the introduction of *<rs:reorientLink>* and *<rs:perceiveLink>*. Table 5 presents the formal context for links.

The formal contexts resulting from both elements and links shown in Tables 4 and 5 provide a systematic representation of how objects are distributed across semantic attributes, establishing a common analytical basis for constructing the concept lattices. The lattices enable the enumeration of all logically valid attribute combinations, the evaluation of the ISO-Space coverage for elements and links, and the identification of required extensions or refinements to represent the full conceptual space of route descriptions in walking context.

Table 5 Formal context for links.

	@event	*@spatialrelation*	*@figure*	*@ground*	*@source*	*@mids*	*@path*	*@trelType*	*@drelType*	*@frameType*	*@mrelType*	*@motion*	*@rs:reorientation*	*@sRelation*	*@rs:perception*	*@mereltype*	*@measure*	*@bound*
<qsLink>	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
<oLink>	0	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
<mLink>	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1
<rs:moveLink>	1	0	1	1	1	1	1	0	0	0	1	1	0	1	0	0	1	0
<rs:reorientLink>	1	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0
<rs:perceiveLink>	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0

4.3 Concept lattice construction

The formal contexts constructed in the previous subsection provide the input for computing the formal concepts for elements and links following standard definitions. Each concept is obtained by computing the closure operation of a set of objects or attributes, resulting in a pair consisting of an extent and an intent. Ordering these formal concepts by extent inclusion produces a complete concept lattice. Each node corresponds to a pairing of an extent and an intent, reflecting how elements cluster according to shared attributes. The concept lattices for elements and links are visualized using Hasse diagrams in Figures 1 and 2.

In the Hasse diagrams (Figures 1 and 2), each node represents a formal concept, with the extent shown in red below the node and the intent in dark blue above it. Concepts absent from ISO-Space are highlighted in purple, those requiring refinement in light yellow, and all others remain uncolored.

- **Elements:** For the *element* formal context, all closed object–attribute pairs were computed and ordered by extent inclusion to obtain the concept lattice shown in Figure 1. This lattice illustrates how elements inherited from ISO-Space, along with newly introduced ones, are organized into a hierarchy based on shared attributes.
- **Links:** For *links*, the formal context is processed analogously: all formal concepts are computed and ordered to form the concept lattice shown in Figure 2. This lattice reveals how existing ISO-Space links and newly proposed links are positioned within the conceptual space based on their attributes.

4.4 Refinements

In the following, the refinements for elements and links are completely described.

- **Elements:** According to Section 3.15 of ISO 24617-7, a *<path>* is defined as *"a location where the focus is on the potential for traversal or which functions as a boundary"*. The specification further notes that *"a <path> might be represented as an undirected graph whose vertices are locations and whose edges signify continuity; i.e., unlike an <eventPath>, a path has no inherent directionality"*. Section 3.3 defines *<eventPath>* as a *"directed path followed by a mover and coincident with a motion-event"*, noting that *"unlike (static) paths such as roads or circular tracks, event-paths are each triggered by a specific motion-event, characterized as being finite directed paths each with a start and an end"*, and further specifies that an *<eventPath>* is a non-consuming tag. A non-consuming tag does not anchor to textual spans, so its *@target* attribute is null. *@target* is an attribute of *<eventPath>* that maintains the textual span.

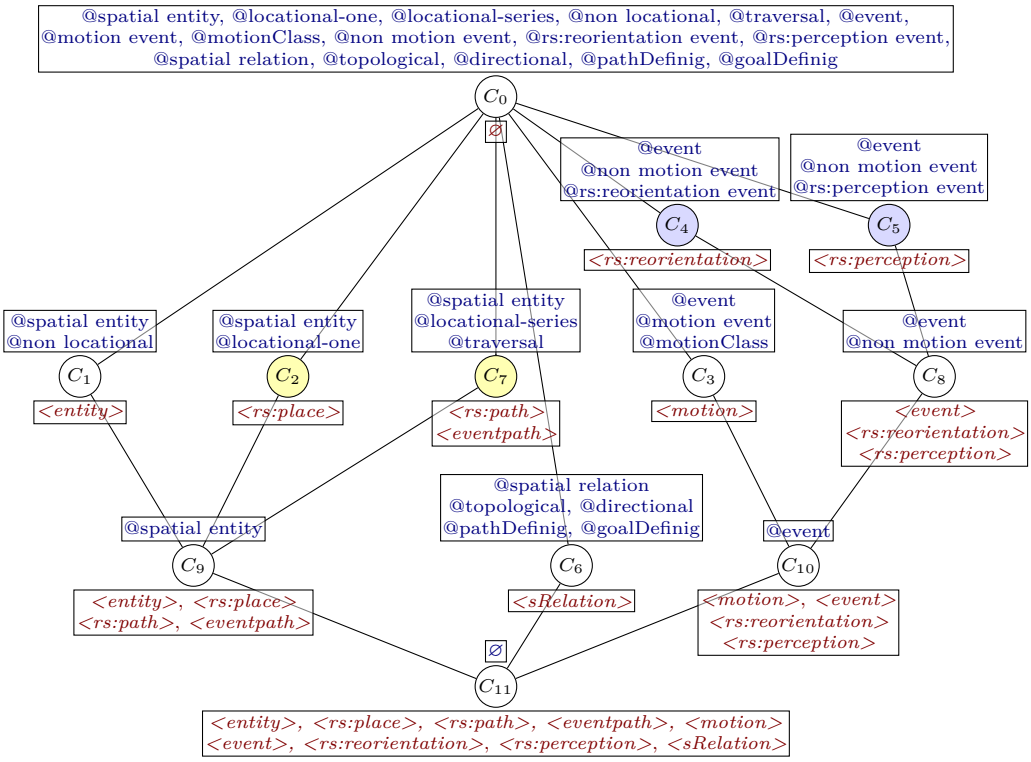


Figure 1 Concept lattice for elements: each node represents a formal concept with an extent (elements below, dark red) and an intent (attributes above, dark blue), ordered by extent inclusion. Nodes absent from ISO-Space are shown in purple, and nodes requiring refinement in yellow.

<*eventPath*> assumes that the path of movement is implicitly recoverable; however, in verbal route descriptions it is often expressed explicitly (e.g., *walk straight along Park Street*). This raises an annotation problem: should an explicit path of movement be annotated as <*path*> or <*eventPath*>? Such paths fit neither the non-consuming definition of <*eventPath*> nor the notion of an undirected <*path*>.

To resolve this disambiguation, either <*eventPath*> must not be non-consuming, or <*path*> must be directional. Representing movement paths via a non-consuming <*eventPath*> complicates annotation and downstream machine-learning tasks; therefore, <*eventPath*> is eliminated and <*path*> is redefined to include direction. Consequently, <*rs:path*> and <*eventPath*> co-occur in node *C7* with identical attributes. Under the refined definition of <*rs:path*>, <*eventPath*> has no distinct attributes, indicating that the two elements are indistinguishable and should be merged.

The next refinement concerns <*place*>. In Section 3.16 of the ISO 24617-7, <*place*> is defined as *"a geographic or administrative entity that is situated at a location"*, annotating *"geographic entities such as lakes and mountains, as well as administrative entities such as towns and counties"*. Under this definition, not all words referring to entities with a shared identification of location can be annotated as <*place*> (e.g., *fountain* in *turn left at the fountain*), limiting its use for place references in route descriptions. In addition, the *@type* attribute of <*place*> is restricted to predefined values and does not allow the annotation of all words that align with the definition in Section 4.1. Therefore, the definition of <*place*> and its *@type* attribute are modified to allow <*rs:place*> to annotate place references in route descriptions.

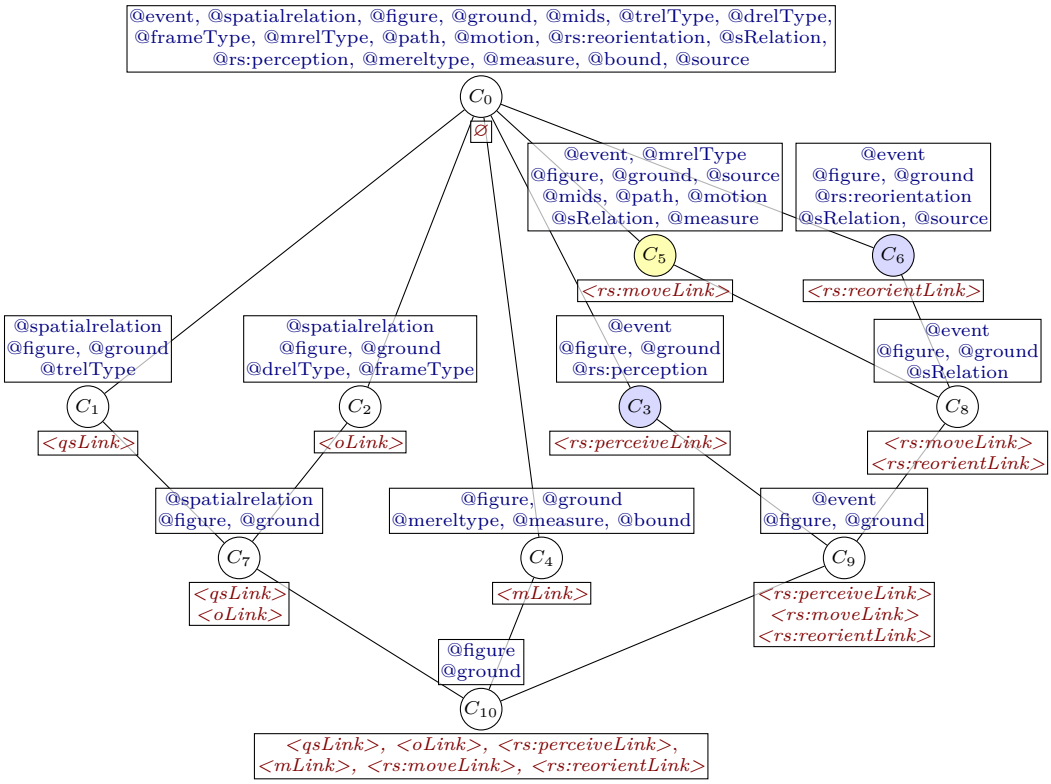


Figure 2 Concept lattice for links: each node represents a formal concept with an extent (links below, dark red) and an intent (attributes above, dark blue), ordered by extent inclusion. Nodes absent from ISO-Space are shown in purple, and nodes requiring refinement in yellow.

- **Links:** According to Section 8.4.3 of ISO 24617-7, *<moveLink>* is defined as a triple $\langle \eta, E, \rho \rangle$, where η is the mover and figure, E is a singleton containing a *<eventPath>* entity serving as the ground, and ρ is the relation between η and E triggered by a *<motion>* event. Its XML attributes include an identifier, a *@relType* (default *traverse* or a *@motionClass*), the figure, the ground, and the trigger. Refinement of *<moveLink>* is motivated by four issues: elimination of *<eventPath>* requires encoding the path of movement, the attributes lack directional information and *<measure>*, and they do not encode the source, mids, or goal of movement. Accordingly, node *C5* in the link concept lattice adds *@sRelation*, *@measure*, *@mids*, *@source*, and *@ground* (the goal of movement), which are absent from the current specification.

4.5 Extensions

Extensions are suggested for situations where concept lattices reveal combinations of attributes not captured by the original specification.

- **Elements:** According to the definition provided in Section 3.2 of the ISO 24617-7 standard, an *<event>* is defined as *"something that can be said to obtain or hold true, to happen, or to occur"*. In the abstract syntax, events constitute the set E, including both motions and non-motions. Section 8.3.6 further clarifies that a non-motion *<event>* corresponds to a TimeML *<event>* that does not involve a change of location.

Section 3.10 defines *<motion>* as an *"action or process involving the translocation of a spatial object, transformation of some spatial property of an object, or change in the conformation of an object"*. In the abstract syntax, motions belong to the set E but are distinguished from non-motions. Section 8.3.5 specifies that a *<motion>* is *"a TimeML event that involves a change of location of some entity"*. When an event is recognized as a motion, *<event>* is replaced by *<motion>* with attributes *@motionType*, *@motionManner*, *@motionGoal*, *@motionClass*, and *@motionSense*, supporting distinctions in manner vs. path, goals, and literal, fictive, or intrinsic interpretations.

Although route descriptions frequently contain reorientation verbs, the definition of *<motion>* excludes them. Reorientation verbs describe a change in facing direction and correspond to Denis's action class of changing orientation [8]. Section 3.10 restricts *<motion>* to events involving translocation, transformation of spatial property, or change in conformation, none of which covers a change in facing direction without displacement or with displacement. Consequently, reorientation verbs fall outside the scope of *<motion>* in ISO-Space and cannot be annotated as such without distorting their semantics. Under the generic definition of *<event>* in Section 3.2, these verbs would be annotated as *<event>*, obscuring their interpretive role; therefore, route descriptions require a distinct tag to capture their semantics.

Perceptual experiences function as descriptive components in route descriptions [2], introducing landmarks along the route or visible from a distance. They participate in the positioning and inspection actions by Denis [8], allowing a moving entity to check orientation based on the visual scene. Under the definition of *<event>* in Section 3.2, these verbs fall under *<event>* but require a dedicated tag distinct from *<event>*.

The concept lattice shows that neither *<reorientation>* nor *<perception>* shares all attributes with existing elements such as *<event>* and *<motion>*, appearing instead in singleton nodes (*C4* and *C5*). As they do not cluster with any ISO-Space elements, they constitute new concepts, leading to an extension of the element taxonomy with *<rs:reorientation>* and *<rs:perception>*. This extension ensures that route instructions involving turns or perceptual cues are explicitly represented and linked through *<rs:reorientLink>* and *<rs:perceiveLink>* to their figure and ground.

- **Links:** The lattice contains singleton concepts (*C3* and *C6*) whose extents consist only of links added to capture attribute combinations, namely *<rs:reorientLink>* and *<rs:perceiveLink>*. These appear as independent nodes because their attributes, *@rs:reorientation* and *@rs:perception*, do not co-occur with any links. Their position reflects that these attribute combinations are conceptually valid yet absent from ISO-Space; therefore, *<rs:reorientLink>* and *<rs:perceiveLink>* complete the conceptual space of links.

5 Integration of refinements and extensions via CASCADES

To answer RQ2, the CASCADES methodology (**C**onceptual analysis, **A**bstract syntax, **S**emantics, and **C**oncrete syntax for **A**nnotation **DES**ign) is used to integrate missing spatial concepts. CASCADES, proposed by Bunt [7], divides annotation design into four stages: 1) *Conceptual Analysis*, which identifies the information to be captured by analyzing relevant linguistic and cognitive theories; 2) *Abstract Syntax*, which defines the vocabulary of basic concepts and their possible combinations as set-theoretic constructs; 3) *Semantics*, which provides a formal interpretation of the abstract syntax; and 4) *Concrete Syntax*, which specifies how annotation structures are represented in a concrete format.

CASCADES provides a procedural framework for translating conceptual analysis into ISO-Space+RP specifications. The refinements and extensions identified in Section 4 are formalized in ISO-Space abstract syntax and XML-based concrete syntax as a machine-readable representation, with DTD declarations provided in Appendix A.

- **Refinements**
  <rs:path>: The attributes of this element remain unchanged from the original standard.
  <rs:place>: The attributes remain unchanged except for *@type*, which is treated as *CDATA* under the modified definition.
  <rs:moveLink>: The attributes include *@id*, *@reltype*, *@figure*, *@ground*, *@source*, *@mids*, *@path*, *@trigger*, *@sRelation*, and *@measure*. The refinements update *@ground* to refer to the goal of movement and add the attributes *@sRelation*, *@measure*, *@source*, *@mids*, and *@path*.
- **Extensions**
  <rs:reorientation> and **<rs:perception>**: The attributes of these elements include *@id* and *@target*.
  <rs:reorientLink>: The attributes include *@id*, *@figure*, *@ground* (the place or path used as a reference for the change in direction), *@source* (the place or path at which the reorientation occurs), *@trigger*, and *@sRelation*.
  <rs:perceiveLink>: The attributes include *@id*, *@figure*, *@ground*, and *@trigger*.

6 Evaluation of the modified annotation scheme

Addressing RQ3, this section presents a systematic evaluation of the ISO-Space+RP scheme for route descriptions. The evaluation assesses whether the ISO-Space+RP scheme improves coverage and consistency and is conducted on the corpus described in Section 3.2. Two annotation rounds are performed, each involving two annotators who are PhD students with research backgrounds relevant to geographic information science, using the INCEpTION annotation tool [14]. Two annotation guidelines are used: the first and second rounds follow ISO-Space and ISO-Space+RP specifications for elements, links, and attributes, respectively.

In the first annotation round, the corpus is annotated using the original ISO-Space specifications and inter-annotator agreement (IAA) is computed; in the second round, it is annotated using the ISO-Space+RP scheme and IAA recalculated. The results are compared using two metrics: *coverage*, defined as the number and proportion of identified elements and links, and *consistency*, measured by Krippendorff's α [17]. Agreement values between 0.41 and 0.60 are interpreted as moderate, values between 0.61 and 0.80 as substantial, and values above 0.81 as almost perfect agreement [18]. Element agreement is computed at the span level, requiring overlapping spans with the same element type, while link agreement applies a strict criterion requiring all link attributes to match or overlap between annotators. The evaluation results, comprising the count and proportion of elements and links (Figure 3) and inter-annotator agreement (Figure 4), with an emphasis on the proposed refinements and extensions, are discussed in the following.

In the first annotation, *<eventPath>* accounts for 11% of the annotated elements and is removed in the ISO-Space+RP scheme without loss of spatial information, as it does not anchor to explicit token spans. Its attributes, *@mids*, *@source*, and *@ground*, are instead represented in *<rs:moveLink>*, ensuring no loss of spatial information. The agreement at the element level is lowest for *<eventPath>* ($\alpha = 0.28$), indicating substantial ambiguity between the two *<eventPath>* and *<path>* discussed in Section 4.4. This ambiguity is further reflected in the counts: although *<eventPath>* is required as the *@ground* of *<moveLink>* in the original scheme, the corpus contains 623 *<moveLink>* but only 403 *<eventPath>*.

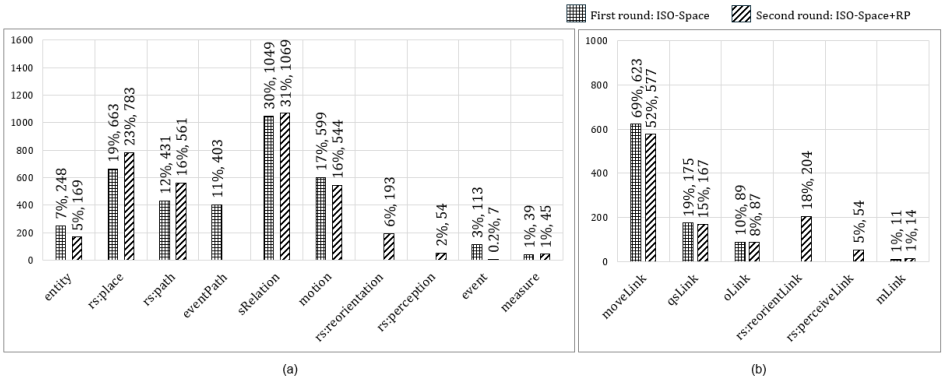


Figure 3 Statistics for two annotation rounds: (a) elements and (b) links. Percentages and counts are shown above bars.

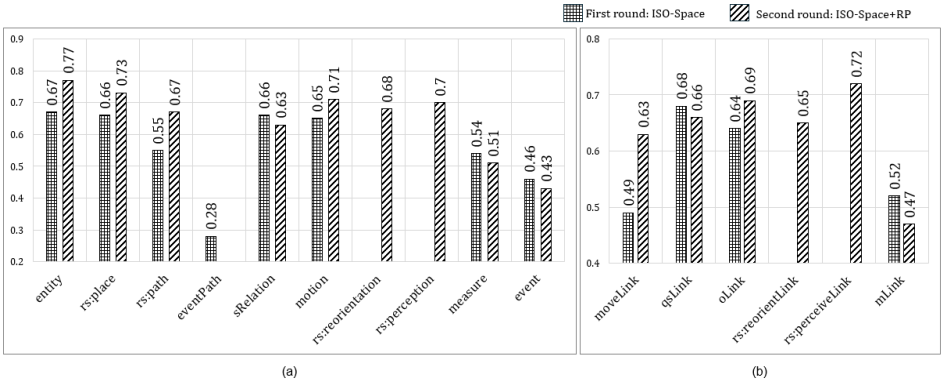


Figure 4 IAA for two annotation rounds: (a) elements and (b) links. Values are shown above each bar.

The number of *<rs:path>* increases from 431 to 561 due to the elimination of *<eventPath>* and the revised definition of *<rs:path>*. In the first annotation, the definition of *<path>* does not allow annotation of explicitly mentioned movement paths, leading to under-annotation, and such references cannot be annotated using the non-consuming *<eventPath>* tag. Under the revised definition, movement paths are annotated as *<rs:path>* in the second annotation. The revised definition and elimination of *<eventPath>* increase the IAA of *<rs:path>* from 0.55 to 0.67 by removing ambiguity in annotating movement path references.

The number of *<rs:place>* annotations increases from 663 to 783. Due to the revised definition of *<rs:place>*, all words that refer to entities with a shared identification of location are annotated as *<rs:place>*. Consequently, *<rs:place>* IAA increases from 0.66 to 0.73.

In ISO-Space scheme, *<moveLink>* includes three attributes, *@figure*, *@motion*, and *@ground*, with *<eventPath>* as the permissible *@ground*. The following statistics demonstrate that modified *<rs:moveLink>* is critical: 29 of 54 *<measure>* (64%), 368 of 561 *<path>* (66%), 556 of 1,069 *<sRelation>* (52%), and 370 of 783 *<place>* (48%) occur within *<rs:moveLink>* but are ignored by ISO-Space scheme. Although *@source*, *@mids*, and *@ground* are attributes of *<eventPath>* in ISO-Space scheme, the high inconsistency of *<eventPath>* ($\alpha = 0.28$) undermines its suitability to hold these attributes. At the link level, *<moveLink>* shows the lowest agreement ($\alpha = 0.49$) in the first round, increasing to $\alpha = 0.63$ in the second round.

In the second annotation, introducing *<rs:perception>* and *<rs:reorientation>* enables 247 elements (7%) and 258 corresponding links, *<rs:perceiveLink>* and *<rs:reorientLink>*, accounting for 23% of all links. Both *<perception>* and *<reorientation>* occur in all four datasets. Across the 12 route descriptions per stratum (urban-walking, natural/rural-walking, indoor-walking, and natural/rural-hiking), counts of *<perception>* are 30, 11, 5, and 54, and counts of *<reorientation>* are 59, 66, 32, and 36, respectively. These results show that the extensions are necessary in walking contexts, as these elements occur consistently in route descriptions regardless of activity type or environmental context. The introduced elements *<rs:perception>* ($\alpha = 0.70$) and *<rs:reorientation>* ($\alpha = 0.68$), and their corresponding links *<rs:perceiveLink>* ($\alpha = 0.72$) and *<rs:reorientLink>* ($\alpha = 0.65$), show substantial agreement.

The total number of elements increases from 3,142 to 3,425, and links from 898 to 1,103. The increases in elements are due to *<rs:reorientation>*, *<rs:perception>*, *<rs:path>*, and *<rs:place>*, while the counts for *<event>*, *<motion>*, and *<entity>* decrease. This decrease does not indicate information loss; instances previously annotated as *<event>* (104 of 113) and *<motion>* (63 of 599) are reclassified as *<rs:reorientation>*, and instances previously annotated as *<entity>* (106 of 248) are reclassified as *<rs:place>*. The increase in link counts results from the introduction of *<rs:perceiveLink>* and *<rs:reorientLink>*, while the number of *<rs:moveLink>* decreases as 63 of 623 instances are reassigned to *<rs:reorientLink>*.

Overall IAA increases from 58% to 67% for elements and from 55% to 66% for links. These gains occur despite reduced agreement for some elements, including *<sRelation>*, *<event>*, and *<measure>*, and for the links *<qsLink>* and *<mLink>*. Refinement of *<rs:place>* and *<rs:path>* and elimination of *<eventPath>* contribute to higher agreement, while restructuring *<rs:moveLink>* also increases IAA despite requiring agreement on more attributes. New introduced elements and links show substantial IAA.

7 Conclusion

Addressing research questions. This paper identifies the spatial elements and links required to represent route descriptions in walikng and hiking contexts. Using Formal Concept Analysis, the study identifies conceptual gaps in ISO-Space relevant to route descriptions, including perceptual expressions and reorientation actions, thus addressing RQ1. These concepts are defined and integrated into the ISO-Space framework through targeted extensions and refinements, using a compatible XML representation that preserves ontological coherence and interoperability, thus addressing RQ2. Two annotation rounds evaluate the impact of the targeted extensions and refinements on inter-annotator agreement and annotation coverage. The results show that ISO-Space+RP scheme increases the annotation coverage without compromising reliability, thus addressing RQ3.

Interpretation and contributions. The increases in element and link counts result from the refinements and extensions and reflect reclassification rather than information loss. The introduction of extensions enables the annotation of *<rs:perception>* and *<rs:reorientation>* and their associated links that occur in route descriptions regardless of the type of activity including walking and hiking or environmental context including urban, natural, rural and indoor. Agreement increases for both elements and links, demonstrating that the refinements improve consistency and that the extensions maintain substantial inter-annotator agreement while remaining compatible with ISO-Space. The ISO-Space+RP scheme enables the training of machine learning models that automatically detect spatial elements and links in walking and hiking contexts beyond those identifiable under ISO-Space. Specifically,

the *<rs:reorientation>* element enables explicit and computational segmentation of scenes, determined by the position and viewing orientation of a moving entity. Segmentation of scenes is required for the interpretation of directional spatial relations. The *<rs:perception>* element introduces visible places that support the interpretation of spatial relations.

Limitations and future work. Some elements and links, including *<measure>*, *<event>*, and *<mLink>*, remain sparse, potentially affecting the stability of the agreement. The evaluation dataset is also restricted to walking and hiking. Route descriptions from other contexts may reveal other conceptual elements and links beyond those proposed in this study. Future work could apply the ISO-Space+RP scheme to broader activity types, add adjudication to resolve disagreements and improve reliability, expand the corpus to better represent under-represented elements and links including *<measure>*, *<event>*, and *<mLink>*, and investigate the impact of ISO-Space+RP scheme on downstream tasks such as route reconstruction and spatial reasoning.

References

1 Niloofar Aflaki, Shaun Russell, and Kristin Stock. Challenges in creating an annotated set of geospatial natural language descriptions (short paper). In *10th International Conference on Geographic Information Science (GIScience 2018)*, pages 20:1–20:6. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.GISCIENCE.2018.20.

2 Gary L Allen. Principles and practices for communicating route knowledge. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 14(4):333–359, 2000.

3 Nicholas Asher and Pierre Sablayrolles. A typology and discourse semantics for motion verbs and spatial PPS in French. *Journal of Semantics*, 12(2):163–209, April 1995. doi:10.1093/jos/12.2.163.

4 Emmon Bach. The algebra of events. *Linguistics and Philosophy*, 9(1):5–16, 1986. doi:10.1007/BF00627432.

5 Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. doi:10.3115/980845.980860.

6 Lamia Belouaer, David Brosset, and Christophe Claramunt. From verbal route descriptions to sketch maps in natural environments. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2996913.2997003.

7 Harry Bunt. A methodology for designing semantic annotations. Technical Report TR 2013-001, TiCC Technical Report, Tilburg Center for Cognition and Communication, Tilburg University, 2013. URL: https://let.uvt.nl/general/people/bunt/docs/CASCADES-TiCC-TR.pdf.

8 Michel Denis. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458, 1997.

9 Michel Denis. *Space and spatial cognition: A multidisciplinary perspective*. Routledge, 2017.

10 Bernhard Ganter, Gerd Stumme, and Rudolf Wille. Formal concept analysis: Theory and applications. *JUCS - Journal of Universal Computer Science*, 10(8):926–926, 2004. doi:10.3217/jucs-010-08.

11 Alexander Henlein, Giuseppe Abrami, Attila Kett, and Alexander Mehler. Transfer of ISO-Space into a 3D environment for annotations and applications. In Harry Bunt, editor, *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 32–35, Marseille, May 2020. European Language Resources Association. URL: https://aclanthology.org/2020.isa-1.4/.

12 Ray Jackendoff. Conceptual semantics and cognitive linguistics. *Cognitive Linguistics*, 7(1):93–129, 1996. doi:10.1515/cogl.1996.7.1.93.

13 Olga Kagan. *The semantics of case*. Key Topics in Semantics and Pragmatics. Cambridge University Press, Cambridge, 2020. doi:10.1017/9781108236867.

14 Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018. The 27th International Conference on Computational Linguistics (COLING 2018). URL: http://tubiblio.ulb.tu-darmstadt.de/106270/.

15 Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. Semeval-2012 task 3: spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 365–373, USA, 2012. Association for Computational Linguistics. URL: https://aclanthology.org/S12-1048/.

16 Parisa Kordjamshidi, Marie-Francine Moens, and Martijn van Otterlo. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 413–420. European Language Resources Association (ELRA), 2010.

17 Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, January 2006. doi:10.1111/j.1468-2958.2004.tb00738.x.

18 J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. URL: http://www.jstor.org/stable/2529310.

19 Kevin Lynch. *The Image of the City*. MIT Press, Cambridge, MA, 1960.

20 Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. SpatialML: Annotation scheme, corpora, and tools. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL: https://aclanthology.org/L08-1017/.

21 Ludovic Moncla, Mauro Gaio, Javier Nogueras-Iso, and Sébastien Mustière. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*, 30(6):1137–1160, 2016. doi:10.1080/13658816.2015.1108422.

22 Philippe Muller. A qualitative theory of motion based on spatio-temporal primitives. In *Principles of Knowledge Representation and Reasoning*, pages 131–143, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

23 Eric Nichols and Fadi Botros. SpRL-CWW: Spatial relation classification with independent multi-class models. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 895–901, Denver, Colorado, June 2015. Association for Computational Linguistics. doi:10.18653/v1/S15-2150.

24 Michał Olek and Maciej Piasecki. Three-stage extraction of spatial relationships using markers. In Ngoc-Than Nguyen, Bogdan Franczyk, André Ludwig, Manuel Nunez, Jan Treur, Gottfried Vossen, and Adrianna Kozierkiewicz, editors, *Advances in Computational Collective Intelligence*, pages 159–172, Cham, 2024. Springer Nature Switzerland.

25 Ross S. Purves, Stephan Winter, and Werner Kuhn. Places in information science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182, October 2019. doi:10.1002/asi.24194.

26 James Pustejovsky, Robert Ingria, Roser Saurí, José M. Castaño, Jessica Littman, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. The specification language TimeML. In Inderjeet Mani, James Pustejovsky, and Robert J. Gaizauskas, editors, *The Language of Time - A Reader*, pages 545–558. Oxford University Press, 2005.

27 James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. SemEval-2015 Task 8: SpaceEval. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June 2015. Association for Computational Linguistics. doi:10.18653/v1/S15-2149.

28 James Pustejovsky, Jessica Moszkowicz, and Marc Verhagen. A linguistically grounded annotation language for spatial information. *Traitement Automatique des Langues*, 53(2):87–113, 2012. URL: https://aclanthology.org/2012.tal-2.4/.

29 James Pustejovsky and Jessica L. Moszkowicz. Integrating motion predicate classes with spatial and temporal annotations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Posters Proceedings, 18-22 August 2008, Manchester, UK*, pages 95–98, 2008. URL: https://aclanthology.org/C08-2024/.

30 Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. BERT-based Spatial Information Extraction. In Parisa Kordjamshidi, Archna Bhatia, Malihe Alikhani, Jason Baldridge, Mohit Bansal, and Marie-Francine Moens, editors, *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17. Association for Computational Linguistics, November 2020. doi:10.18653/v1/2020.splu-1.2.

31 L. Talmy. Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen, editor, *Language Typology and Syntactic Description. Volume III: Grammatical Categories and the Lexicon*. Cambridge University Press, 1985.

32 Thora Tenbrink. Relevance in spatial navigation and communication. In *International Conference on Spatial Cognition*, pages 358–377. Springer, 2012. doi:10.1007/978-3-642-32732-2_23.

33 Barbara Tversky and Paul U. Lee. How space structures language. In Christian Freksa, Christopher Habel, and Karl F. Wender, editors, *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pages 157–175. Springer, Berlin, Heidelberg, 1998. doi:10.1007/3-540-69342-4_8.

34 Maria Vasardani, Sabine Timpf, Stephan Winter, and Martin Tomko. From descriptions to depictions: A conceptual framework. In Thora Tenbrink, John Stell, Antony Galton, and Zena Wood, editors, *Spatial Information Theory*, pages 299–319, Cham, 2013. Springer International Publishing. doi:10.1007/978-3-319-01790-7_17.

35 Feng Wang, Peifeng Li, and Qiaoming Zhu. A Hybrid Model of Classification and Generation for Spatial Relation Extraction. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1915–1924, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL: https://aclanthology.org/2022.coling-1.166/.

36 Rudolf Wille. Concept lattices and conceptual knowledge systems. *Computers & Mathematics with Applications*, 23(6):493–515, 1992. doi:10.1016/0898-1221(92)90120-7.

37 Stephan Winter, Ehsan Hamzei, Nico Van de Weghe, and Kristien Ooms. A Graph Representation for Verbal Indoor Route Descriptions. In Sarah Creem-Regehr, Johannes Schöning, and Alexander Klippel, editors, *Spatial Cognition XI*, pages 77–91, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-319-96385-3_6.

38 Jordan Zlatev. Spatial Semantics. In *The Oxford Handbook of Cognitive Linguistics*. Oxford University Press, 2010. doi:10.1093/oxfordhb/9780199738632.013.0013.

A Appendix: Document Type Definition of refinements and extensions

Document Type Definition (DTD) of refinements and extensions are provided below.

Refinements

<rs:moveLink>

Attributes for <rs:moveLink>

<!ELEMENT rs:moveLink EMPTY >
<!ATTLIST rs:moveLink id ID prefix="mvL" #REQUIRED >
<!ATTLIST rs:moveLink relType (traverse | CDATA) #REQUIRED >
<!-- relation type of motion, defaults to traverse or motion class -->
<!ATTLIST rs:moveLink figure IDREF #REQUIRED >
<!-- identifier of the entity that moves along the path -->
<!ATTLIST rs:moveLink ground IDREF #REQUIRED >
<!-- identifier of the relevant place or path -->
<!ATTLIST rs:moveLink source IDREF #REQUIRED >
<!-- identifier of the relevant place or path -->
<!ATTLIST rs:moveLink mids IDREF #REQUIRED >
<!-- identifier of the relevant place or path -->
<!ATTLIST rs:moveLink path IDREF #REQUIRED >
<!-- identifier of the path of movement -->
<!ATTLIST rs:moveLink trigger IDREF #IMPLIED >
<!-- identifier of the motion event that triggered the link -->
<!ATTLIST rs:moveLink sRelation IDREF #REQUIRED >
<!-- identifier of sRelation related to motion event -->
<!ATTLIST rs:moveLink measure IDREF #REQUIRED >
<!-- identifier of the corresponding measure -->
<!ATTLIST rs:moveLink comment CDATA #IMPLIED >

▶ **Example 1.** I_{x1} $walk_{m1}$ to_{sr1} the $fountain_{pl1}$ for $[2\ minutes]_{me1}$.

<entity xml:id ="x1" target="I"/>
<motion xml:id="m1" target="walk" motionClass="moveInternal" motionSense="literal"/>
<sRelation xml:id="sr1" target="to" type="goalDefining"/>
<place xml:id="pl1" target="fountain"/>
<measure xml:id="me1" target="2 minutes" value="2" unit="minutes"/>
<rs:moveLink xml:id="mvL1" relType="traverse" figure="#x1" ground="#pl1"
trigger="#m1" sRelation="#sr1" measure="#me1"/>

Extensions

<rs:reorientation>

Attributes for <rs:reorientation>

<!ELEMENT rs:reorientation (#PCDATA) >
<!ATTLIST rs:reorientation id ID prefix="r" #REQUIRED >
<!ATTLIST rs:reorientation target (IDREF | CDATA) #REQUIRED >
<!ATTLIST rs:reorientation comment CDATA #IMPLIED >

▶ **Example 2.** I turn_{r1} left into Bourke street.

<rs:reorientation xml:id="r1" target="turn"/>

<rs:perception>

Attributes for <rs:perception>

<!ELEMENT rs:perception (#PCDATA) >
<!ATTLIST rs:perception id ID prefix="r" #REQUIRED >
<!ATTLIST rs:perception target (IDREF | CDATA) #REQUIRED >
<!ATTLIST rs:perception comment CDATA #IMPLIED >

▶ **Example 3.** You see_{pe1} the library.

<rs:perception xml:id="pe1" target="see"/>

<rs:reorientLink>

Attributes for <rs:reorientLink>

<!ELEMENT rs:reorientLink EMPTY >
<!ATTLIST rs:reorientLink id ID prefix="reL" #REQUIRED >
<!ATTLIST rs:reorientLink figure IDREF #REQUIRED >
<!-- identifier of the entity whose orientation changes -->
<!ATTLIST rs:reorientLink ground IDREF #REQUIRED >
<!-- identifier of the referenced place or path -->
<!ATTLIST rs:reorientLink source IDREF #REQUIRED >
<!-- identifier of the referenced place or path -->
<!ATTLIST rs:reorientLink trigger IDREF #IMPLIED >
<!-- identifier of the reorientation event that triggered the link -->
<!ATTLIST rs:reorientLink sRelation IDREF #REQUIRED >
<!-- identifier of sRelation related to reorientation event -->
<!ATTLIST rs:reorientLink comment CDATA #IMPLIED >

▶ **Example 4.** I_{x1} turn_{r1} left_{sr1} into_{sr2} $[\text{Bourke street}]_{\text{p1}}$.

<entity xml:id="x1" target="I"/>
<rs:reorientation xml:id="r1" target="turn"/>
<sRelation xml:id="sr1" target="left"/>
<sRelation xml:id="sr2" target="into"/>
<path xml:id="p1" target="Bourke Street"/>
<rs:reorientLink xml:id="reL1" figure="#x1" ground="#p1" trigger="#r1"
sRelation="#sr1, #sr2"/>

<rs:perceiveLink>

Attributes for <rs:perceiveLink>

<!ELEMENT rs:perceiveLink EMPTY >
<!ATTLIST rs:perceiveLink id ID prefix="peL" #REQUIRED >
<!ATTLIST rs:perceiveLink figure IDREF #REQUIRED >
<!-- identifier of the perceiving entity -->
<!ATTLIST rs:perceiveLink ground IDREF #REQUIRED >
<!-- identifier of the perceived place or path -->
<!ATTLIST rs:perceiveLink trigger IDREF #REQUIRED >
<!-- identifier of the perception event that triggered the link -->
<!ATTLIST rs:perceiveLink comment CDATA #IMPLIED >

▶ **Example 5.** You_{x1} see_{pe1} the $\text{library}_{\text{pl1}}$.

```
<entity xml:id="x1" target="You"/>
<rs:perception xml:id="pe1" target="see"/>
<place xml:id="pl1" target="library"/>
<rs:perceiveLink xml:id="peL1" figure="#x1" ground="#pl1" trigger="#pe1"/>
```

Mapinou: A Game to Assess the Skill of Using Pan-Scalar Maps

Justin Berli
Univ Gustave Eiffel, ENSG, IGN, LASTIG, F-77420 Champs-sur-Marne, France

Guillaume Touya
Univ Gustave Eiffel, ENSG, IGN, LASTIG, F-77420 Champs-sur-Marne, France

Abstract

The widespread adoption of topographic maps on smartphones, typically implemented as pan-scalar digital cartographic interfaces, has altered the competencies required for their effective interpretation and use. The skills necessary to navigate and analyze these maps now differ substantially from those associated with reading and employing traditional paper-based topographic maps. To facilitate a more rigorous assessment of these emerging competencies, the paper introduces a serious game, termed Mapinou, which is designed to be executable on any modern smartphone device. The game alternates between tasks that require the use of a pan-scalar map and spatial skill tests to study potential relations between the spatial skills of users and their competencies in pan-scalar map use. Our goal when designing this game was to propose a predictor of pan-scalar map use skills, and the spatial skill tests embedded in the current version of the game are intended to test the robustness of this prediction. In this paper, we analyze the data collected from almost two hundred early players of the game. The results show a weak but significant relation between the success in the game and the success to the water-level task, which is known to be related to spatial skills.

2012 ACM Subject Classification Applied computing → Cartography

Keywords and phrases cartography, pan-scalar map, sense of direction, spatial skill, serious game

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.13

Supplementary Material *Software*: `https://github.com/LostInZoom/mapinou/`
archived at `swh:1:dir:b485063356204f36b4a6208d2753ce9b615ccf6f`

Funding This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101003012, LostInZoom).

Acknowledgements We would like to thank all the anonymous players of Mapinou, so far and in the future, for their contribution to science. We would also like to thank Paul Bourcier, Maïeul Gruget, Bérénice Le Mao, Quentin Potié, and Laura Wenclik for their useful suggestions during the design of the game.

1 Introduction

Reading a topographic map used to be a niche activity that required the unfolding of a large piece of paper, and decoding a map with a high density of information. Not everyone was able to properly read a topographic map. Now that almost everyone has access to topographic maps of the world at any scale, thanks to their smartphone, and use them for daily navigation tasks, it is important to wonder if people have the skills to use these new topographic maps. Here, it should be noted that we consider Google Maps, Apple Plan, or even the default OpenStreetMap map as those new forms of topographic maps, even if they do not represent contour lines and relief as detailed as 1:25,000 scale maps from Swiss Topo would. Those maps are different from the paper, single-scaled topographic maps, as they are pan-scalar, i.e., zoomable, interactive, and multi-scale [14]. The main differences are

© Justin Berli and Guillaume Touya;
licensed under Creative Commons License CC-BY 4.0

twofold: the size of a display with a small screen for pan-scalar maps compared to the large piece of paper for printed maps; and the interactivity possible with the pan-scalar maps, through touch gestures. As a consequence, the skills required to read those maps might be different from the skills required to read topographic paper maps. For instance, the users of pan-scalar maps may be subject to cartographic disorientation [43] and to the desert fog effect (i.e., the disappearance of the known landmarks that were still visible in the previous view [20]), when they pan and zoom in the map [42]. From our observations, some users are more subject to the desert fog and to disorientation than others. More generally, some users are more crafty than others with the use of hand gestures to pan and zoom in a pan-scalar map [45], as the pinch gesture can be complex to achieve with precision [44, 31].

The skills and competencies required to read a topographic map have been studied by researchers. Those skills include specific competencies such as reading contour lines, or using a scale bar [41, 2], but also include cognitive processes that are not specific to map reading (e.g. mental rotation, distance estimation, symbol identification) [28, 46, 26]. Some of these general cognitive skills can be summarized as the visual-spatial ability, which can be defined as the ability to visualize and manipulate geometrical figures in space [46]. But do those competencies translate to the use of pan-scalar maps on smartphones? Using and understanding pan-scalar maps requires visual-spatial abilities, but as mentioned earlier, it also requires at least touch gesture skills and the mental ability to relate scales to one another. We are convinced that it is important to better understand the competencies required to use this new medium. As a first step, we want to be able to discriminate between the users who succeed in using these maps and those who fail, or at least those who face difficulties.

To answer such questions, we need to collect data from a very large and diverse sample of the population. Gamification constitutes a strategy for fostering participation in citizen science by incorporating game-like elements and mechanics into research activities [13, 27]. In recent years, gamification has been increasingly integrated into platforms and projects related to volunteered geographic information (VGI) [1, 39, 5]. Beyond the mere collection of geospatial data, serious geo-games [37] can be designed to serve a variety of additional purposes. They can function as educational tools for teaching geography and cultivating spatial competencies, and they can also be employed as experimental instruments in the empirical sciences [38]. Furthermore, some geo-games are developed primarily for recreational use, as evidenced by the widespread popularity of applications such as Geocaching [30] and GeoGuessr.

Our principal source of inspiration for the design of a serious game targeting map-use skills is Sea Hero Quest, a game co-developed by academic researchers and a video-game studio [40]. This mobile game was specifically created to investigate demographic factors influencing spatial skills and their decline, and it reached a large user base of 3.9 million players. The resulting dataset has made it possible to identify global determinants of spatial abilities, including the geographic region in which individuals were born and raised, as well as gender-related differences [8, 7, 6]. Although our ambitions are considerably more modest – particularly because our serious game has been developed without the collaboration of a professional video-game studio – we nevertheless aim to recruit several thousand players. This will enable us to gain a more comprehensive understanding of pan-scalar map-use skills.

In this paper, we introduce Mapinou, a serious, mobile-based geo-game that we have developed. The game entails intensive engagement with and exploration of a topographic, pan-scalar map. Within Mapinou, map-use tasks are interleaved with standardized spatial ability tests drawn from the literature, thereby enabling direct comparison between players' in-game performance and their outcomes on established spatial skill measures. Our primary

objective with Mapinou is to use the game as a predictor of pan-scalar map reading and using skills. If there is light or strong correlations between the game and these 4 spatial skill tests, the game could be used as a predictor, as pan-scalar map reading and using skills are partly related to spatial skills. A futher objective is to better understand of the global determinants of these pan-scalar map using skills (e.g., age, gender, prior cartographic experience). In this article, we report the initial empirical findings derived from data collected with the first cohort of players.

2 Materials and Methods

2.1 Experimental Design

In this study, we are interested in measuring the skills of users in their use of pan-scalar maps, so we opted for a between-subjects design. Each user of Mapinou plays the same game, with the same levels in the same order, and also performs the same spatial skill tests.

The game consists in 16 levels of increasing difficulty. Each level requires solving two distinct tasks. The game is not designed to be completed in a single continuous session. Each level, with a duration ranging from one to five minutes, can be undertaken in a separate session, as the player's progress is persistently stored.

A map is a multi-purpose communication tool, and it can be used for achieving a wide variety of tasks [9, 4, 22]. The interactivity of the pan-scalar maps adds new usages of the map compared to printed maps [35]. If we focus on the most frequent tasks map users achieve with pan-scalar maps, we can cite *global search*, i.e., searching the whole map area for specific targets or locations, *distance estimation*, *route planning*, and *route following* [21, 33]. The task that we designed for each level of the game combines these four generic tasks, and is composed of two subtasks.

The first task consists in *finding a target*: the rabbit the user plays in the game gives textual instructions about search areas. For instance, the rabbit says "I must be in the south of Paris" (Figure 1(a)). Then, if the user zooms in on the south of Paris, another instruction is triggered. Instructions are chained until the user has zoomed enough to visualise and point the target (usually a small town) in the map. When the user is lost, the rabbit gives messages to find the right track (usually by zooming out). The score in this task increments as seconds pass. The goal is to find the target as quickly as possible. There is no penalty for pointing at the wrong target; this only triggers a message "I am not there". This task is mainly a *global search* task.

The second task mixes *route planning*, *distance estimation*, and *route following*, as the user is asked to find a route to reach a rabbit friend located further on the map (Figure 1(b)). The user has to point a location on the map and then the rabbit runs to that location using the roads and the paths from the map. Similar to the first task, the score increments as seconds pass. Pointing to a location to go to is only possible at large scales, which forces the user to memorize the route or to zoom in and out to adjust the route as the rabbit moves. To make the task harder, and force the user to read the map and zoom in and out frequently, we added obstacles to avoid, in the form of enemies: snakes, eagles, and hunters. Those enemies have a radius of influence, and if the rabbit runs into those areas, the score increments with a penalty (Figure 1(c)). To add another layer of reasoning, we introduced vegetables that appear when the rabbit comes close. Eating vegetables gives a bonus to the score, so the user is really incentivized to regularly update the route.

Alternating with the pan-scalar map-use tasks described above, Mapinou presents four established assessments of spatial ability. Figure 2 illustrates how these four tests are embedded within the game environment. First, participants complete a digital implementation

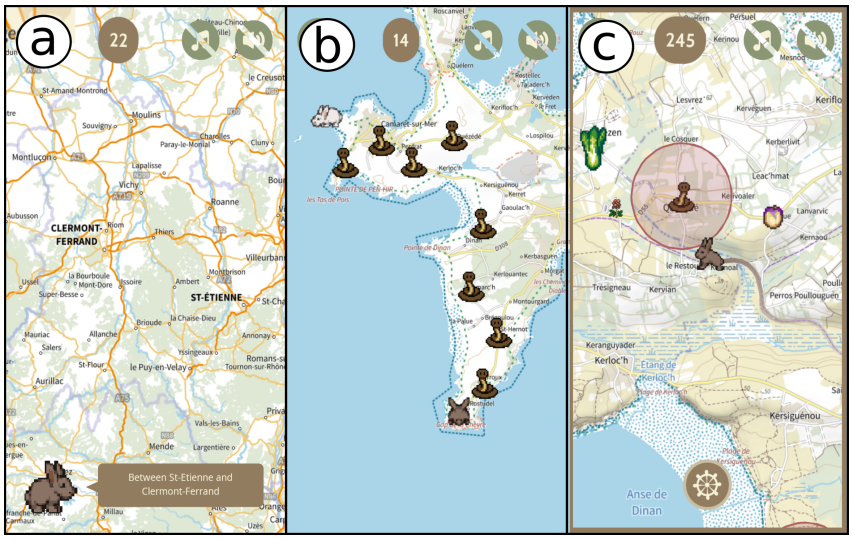


Figure 1 Two tasks to perform during a Mapinou game: (a) find the rabbit using the textual indications; (b) the route planning task; (c) enemies and rewards appear during the second task. All three captions are extracted from the first games available in Mapinou.

of the water-level task. Originally introduced by Piaget [32], this task depicts a cross-sectional view of a bottle filled halfway with water. In the test, a rotated version of the same bottle is shown, and the participant must draw the line corresponding to the interface between water and air. The correct response is the horizontal line that divides the interior of the bottle into two regions of equal area. This task has been extensively employed in prior research on spatial ability, as it has been shown to correlate strongly with other spatial cognition measures [24, 49, 25]. In our application, the line is drawn by tapping twice on the touchscreen to specify the start and end points of the line segment. Because of the inherent imprecision of touch input, we do not require a perfectly horizontal line for a response to be considered correct. Each participant is presented with five bottles, each rotated at a different angle.

We then used the Santa Barbara Sense of Direction Scale [17] that is composed of 15 questions about the self-declared sense of direction of participants. For each assertion (e.g., "I am very good at giving directions"), the participant decides on a Likert scale with 7 values if they strongly agree or strongly disagree with the assertion. This scale was used to study the global determinants of spatial ability [50], to better understand spatial ability [18, 16], or just as a proxy measure of spatial ability to characterise the sample of participants in user studies.

The third spatial ability test is the Perspective Taking Spatial Orientation Test (PTSOT) [23, 18]. In this test, several distinctive symbols are drawn on a paper or displayed on a screen, and the participant is given several questions. For each question, the participant should imagine they are standing at the location of one these symbols (which will be named at the center of the circle), and they are facing at a second of these symbols, named at the top of the circle; they are asked to draw an arrow from the center object showing the direction to a third symbol from this facing orientation. This test requires perspective-taking and mental rotations as cognitive processes. In the application, we used a recent revised version of the PTSOT [11], and we changed the symbols for the sprites used in the game (a rabbit, a snake, an eagle, a pigeon, a zucchini, a hunter). The user has to point to the location of the target symbol on the circle for the line to be drawn.

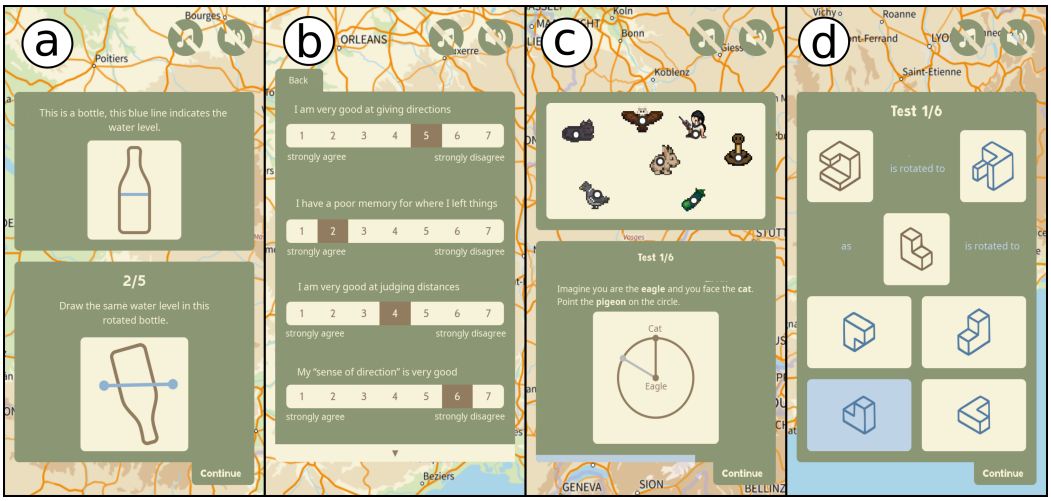


Figure 2 Screen capture of the four spatial skill tests included in the game: (a) the water-level task; (b) the Santa Barbara Sense of Direction Scale; (c) the Spatial Orientation Test (PTSOT); (d) Purdue spatial visualization test for rotations.

Table 1 Number of Mapinou users according to the level they reached in the game.

level reached	level #1	level #4	level #8	level #12	level #16
nb of users	196	112	65	46	35

The final spatial ability test we used is the Purdue Spatial Visualization Test for rotations [15]. The test consists of thirty questions of increasing difficulty related to mental rotation, where the participants are asked to achieve mental rotations of 3D volumes and select the correct rotation among several propositions. Mental rotation tests are often used to measure spatial ability, and other similar tests exist [48]. In the Mapinou application, we selected 5 of the 30 questions from the original test, with varying difficulty, as the full test is expected to last 20 minutes, which would be too long for inclusion in the game.

2.2 Participants

Considering our end goals with this game, we want to reach as many users as possible, encompassing all existing profiles in the population. We started with a small-scale recruitment process, located in our university, where we put some posters on the walls. We also participated in several events of science outreach within the university. A button to share the game on social networks is included in the application, and we expected a minor recruitment process through the social networks of the players. At the time we started the data analysis for this paper, 958 players had started the game, for 1902 games played.

We then filtered out most of these players. First, we removed all the players who were identified as playing on a desktop computer, using the information given by their web browser. We also removed all the players who installed the game on their mobile phone but did not play a game. Table 1 shows the number of players in the data collected for this paper, the level they reached in their game completion. Reaching level 12 means that the player played at least 12 games. In total, 1250 games have been played in this dataset.

Table 2 shows the age distribution of the 196 players. Most of the players are situated between 18 and 44, with a quite even distribution between the three corresponding age categories. For now, the sample is too small in the other age categories to analyze the relation between age and pan-scalar map-using skills.

Table 2 Age distribution of the users in this study.

age	–18	18–24	25–34	35–44	45–59	60+	no answer
count	4	50	58	45	15	8	16

Table 3 shows the gender distribution for the participants, where "other" stands for different possible gender answers, including "no answer". The distribution between male and female players is fairly even.

Table 3 Gender distribution of the users in this study.

gender	female	male	other
count	82	94	20

Finally, Table 4 shows the self-declared map usage frequency categories among our 196 participants. Most of the participants regularly use pan-scalar maps on mobile devices, either everyday (85 participants) or several times a week (82 participants). Given the usage figures for applications such as Google Maps [36] or Apple Plan, it is not surprising to have more players in these two categories.

Table 4 Map usage distribution of the users in this study.

Map usage	everyday	several times a week	several times a month	almost never	no answer
count	85	82	16	1	12

2.3 Materials

In Mapinou, each level requires the exploration of a map, and we did not create this pan-scalar map; we decided to reuse an existing one. We used a pan-scalar map called Plan IGN, designed by the French National Mapping Agency. It is the default map available on the French administration cartographic platform[1], and it is freely available through APIs. In terms of style and content, Plan IGN is quite similar to Google Maps and Apple Plan.

We developed a game-construction application in which we defined 16 distinct games set in heterogeneous regions and landscapes. These 16 games are organized according to progressively increasing difficulty. In phase 1, the difficulty is determined by the degree of precision of the navigational indications provided to the user. For instance, an easier game would indicate "I am between Lyon and Saint-Etienne", while a harder game would just indicate "I am east of Lyon". In phase 2, difficulty is modulated by the number and spatial distribution of enemies, as well as by the presence and configuration of natural obstacles between the starting and ending points of the route (e.g., lakes, large rivers, mountain ranges). As one game is composed of phase 1 then phase 2, we modulated the difficulty of each phase to obtain a game difficulty that is slowly increasing as new games are reached.

[1] `https://cartes.gouv.fr/`

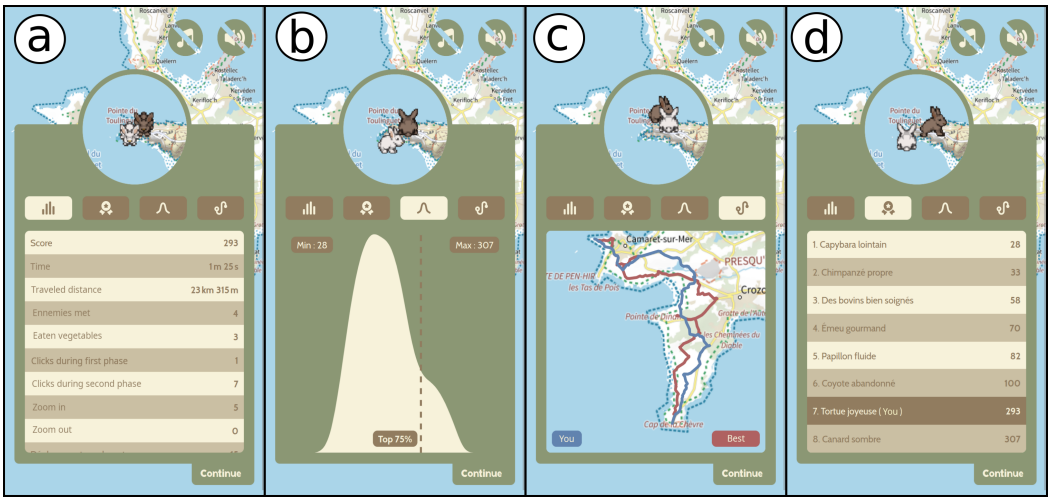


Figure 3 4 screen captures of the game visualizations after one level is finished: (a) summarization of the measures describing the run; (b) the position of the player score in the distribution of this specific level; (c) the comparison of the route of the player compared to the best score route; (d) the score ranking.

The game application was designed with open Javascript libraries, including Maplibre for the cartographic visualizations. The routes chosen when the rabbit moves in the game are based on the routing API from the French National Mapping Agency. The code of the application is available openly on Github[2].

We need the players to play the game several times, as we do not expect the game to be finished in a single session. As a result, we introduced several gamification features to push players into further games after the first session. For instance, Figure 3 shows several features appearing after a level is completed: the global score ranking, the visualization of the route compared to the best score's route, global statistics, and the position of the player in the distribution of scores. Completing some levels also unlocks new skins for the rabbit played in the game.

In-game auditory effects and an original musical score specifically composed by the authors are incorporated to enhance players' sense of immersion within the game environment.

2.4 Apparatus

The game is meant to be played on smartphones, and it is compatible with most of the devices currently in use. Users are invited to install the application on their personal smartphones, using a QR code or a link. As explained in the previous subsection, the second task of each level relies on a routing web API to compute the shortest path from the position of the rabbit to the pointed position. Without a stable internet connection, the displacement of the rabbit can be very slow. As a consequence, the game requires a stable connection and we prevent starting a trial/level otherwise.

Given our focus on the use of pan-scalar maps on mobile phones, we record the screen size of the device used to play the game and exclude devices whose screens fall outside a predefined acceptable size range, either being too small or excessively large. For now, no exclusion of participants has been performed due to screen size, as the data collected from the web browser already identifies desktop users.

[2] `https://github.com/LostInZoom/mapinou/`

Figure 4 Screen capture of the view when the user starts level 1 of regular games, after the water-level task.

2.5 Procedure

After installing the application on their mobile phone, users are first ask to give informed consent. Then, they are asked to fill a short demographic questionnaire with questions about age, gender and map usage. Then, as the game mechanisms can be complex, we introduce the two tasks of each level with a tutorial, where the users can try all the in-game interactions with the map and with the rabbits. In this tutorial, the players are told that the aim is to perform each phase/task as quickly as possible.

Then, the player has to complete 4 levels composed of 4 games that count as trials in our experiment (Figure 4). Completing one game of a level unlocks the following game, and completing the four games of a level unlocks the next level. Between each level, the user has to complete one of the four spatial ability tests. Before level 1, the user starts with the water-level task, then completes the Santa Barbara Sense of Direction Scale between levels 1 and 2, then completes the PTSOT between levels 2 and 3, and finally completes the Purdue spatial visualization test for rotations between levels 3 and 4. The game was presented in French, and all participants are French speakers, though not necessarily native speakers. Within a level, the four games do not have the exact same difficulty, even if it is rather close as we tried to keep an increasing difficulty between game 1 and game 16.

2.6 Statistical Analysis

As mentioned earlier, in this first study based on the data collected with Mapinou, we only focus on the water-level task and the quality of the results in the game, because there are not enough players who reached the levels where the other spatial skill tests are performed. The analysis in relation to these tests will be conducted in further studies.

Because the games do not present the same difficulty, the score ranges can vary widely between games, making it difficult to compare scores or use the mean to assess a player's performance. We also noticed that our expected difficulty was not matching the actual difficulty for players: in some games that we considered difficult, many players found hacks and managed to obtain very low (good) scores, while some games that were considered as easier found the players facing more difficulty than expected. As a result, we decided to

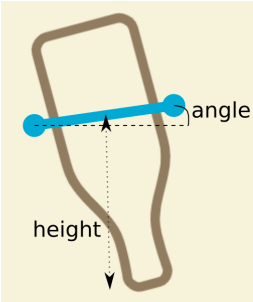


Figure 5 Illustration of the method used to score the drawing from the water-level task.

rather use the z score instead of the raw score obtained by players. The z score can be computed with the following equation, and is a normalization of the score based on the standard deviation of scores in one specific game:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the raw score, μ is the mean of the scores in a level, and σ is the standard deviation of the scores in this level.

Regarding the water-level task, we summarized the 5 drawings with a score from 0 to 5: for each drawing, if the angle between the drawn line and the horizontal is below 10° and if the height of the drawn line is strictly between 30% and 70% of the bottle, the player scores 1 for that drawing (Figure 5); otherwise, they score 0 for that drawing; the total score is the sum of the scores for the five drawings.

Subsequently, we conduct two types of statistical analyses on the results. First, when comparing two ordinal variables, for example the water-level score and the z-score, we assess whether a relationship exists between their distributions. Because none of our distributions meets the assumption of normality, we employ Spearman's rank correlation test to evaluate these regressions.

When comparing a score with categorical data (e.g. the self-declared map usage category, or the age category), we use Kruskal-Wallis tests to check if the categories influence the score.

Finally, we want to use Mapinou in the future as a short test procedure to characterize the sample population in user studies, but it would be too long to have those users play 16 games. So we will compute the minimal sample of games to play to have a mean score representative of the potential final mean score. First, we have to compute the volatility of the scores for the players that played many games (we consider a threshold of 8 games):

$$volatility = \frac{\sum_{i=1}^{k} \sigma_i}{E} \tag{2}$$

where k is the number of players who played more than 8 games, and σ_i is the standard deviation of z scores of the i^{th} player with more than 8 games played.

From the volatility, we can compute n as the minimum number of games to play to use the score mean with a 95% confidence:

$$n = \frac{1.96 \times volatility}{E} \tag{3}$$

where E is the expected error margin on the z score.

Table 5 Distribution of the scores of the users at the water-level task.

score	0	1	2	3	4	5
count	43	4	8	22	30	89

3 Results

Table 5 presents the first set of results, namely the distribution of water-level scores for the 196 participants in our sample. As expected, the most frequent score is 5, corresponding to five correct responses, obtained by 89 participants. An additional 30 participants obtained a score of 4. This latter result can plausibly be attributed to the difficulty of drawing a perfectly horizontal line on a touch screen. Given that our correctness criterion was a strict angular deviation threshold of 10°, several participants who might have produced correct responses using traditional pencil-and-paper methods did not reach the same level of accuracy when using the mobile application. 22% of the participants obtained a score of 0, which is plausible if we consider the failure rates obtained in past studies [49, 24], but looking at the actual answers from some of these players, we have to consider that some players just sidestepped the task by drawing quickly random lines (we recorded the time spent on each question of the water-level task). We did not find a way to automatically identify all those players, as the difference is too thin between a quick player and a player that sidestepped the task. It means that we have to be careful when analyzing the score of the players with this score.

Figure 6 shows the distribution of the mean z score of the players by the water-level score. Apart from the players with a score of zero that are scattered between good (low) z-scores and bad (high) z-scores, this plot shows a tendency of correlation: players with higher water-level scores have better/lower z-scores in the game. The Spearman's correlation is -0.185 ($p - value = 9.2667e - 03$) which confirms a weak but significant correlation. If we exclude the players who scored zero at the water-level task, the Spearman's correlation is stronger and still significant (-0.216; $p - value = 7.4507e - 03$). This result is a first indication that Mapinou might be used as a predictor for pan-scalar map use competency.

Figure 7 shows the box plot of the z-scores by age category. We ran a Kruskal-Wallis test only on the three most populated age categories, namely "18-24", "25-34" and "35-44", which is significant ($p - value : 2.6685e - 02$). we ran a post-hoc Dunn test with a Bonferroni correction, and the category "35-44" is significantly different from "25-34" ($p - value : 0.044$), but the other pairs are not significantly different.

Figure 8 shows the box plot of the z-scores by gender category. As we do not expect any homogeneity in the other category, it was not tested further. We ran a Kruskal-Wallis test on the other two gender categories (H = 6.191; $p - value : 1.2837 \cdot 10^{-2}$), which shows that male players are performing better than female players in the game.

The third demographic factor considered is map usage. Figure 9 presents the boxplot of the z-scores stratified by map usage category. As only the first two categories (players who use maps every day or several times per week) include a sufficient number of participants for statistical analysis, we conducted a Kruskal–Wallis test restricted to these two groups (H = 5.313; $p = 2.1172 \cdot 10^{-2}$). The results indicate that players who use maps on a daily basis achieve significantly higher performance scores than those who use maps only several times per week.

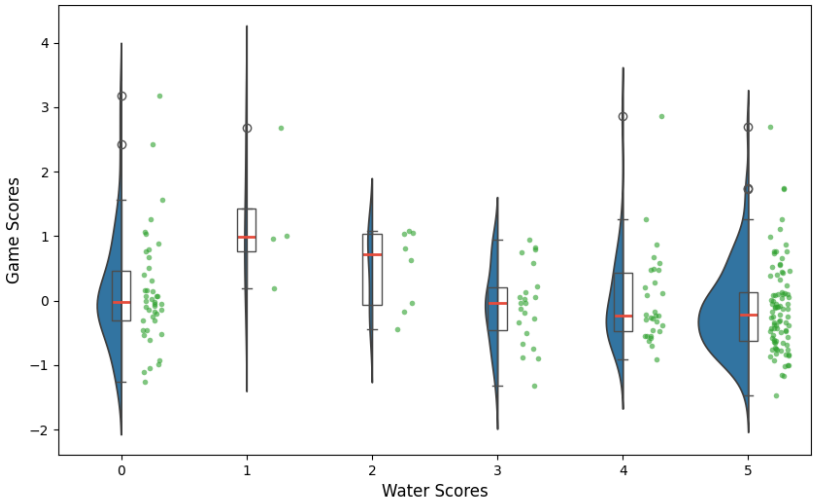


Figure 6 Raincloud plot of the z-score of the participants by water-level score. A raincloud plot combines a box plot, a scatter plot, and a violin plot.

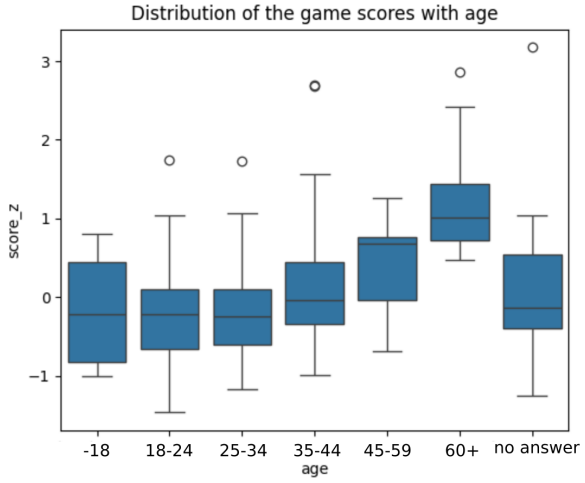


Figure 7 Box plot of the distribution of game z-scores by declared age categories.

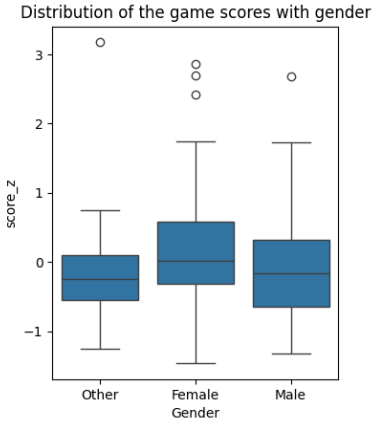


Figure 8 Box plot of the distribution of game z-scores by declared gender. All none "Male" or "Female" answers are aggregated to "Other" in this plot.

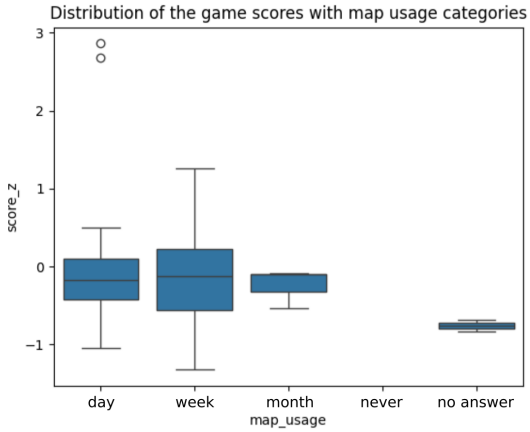


Figure 9 Box plot of the distribution of game z-scores by declared map usage. Day means several times a day, week means several times a week, et cetera.

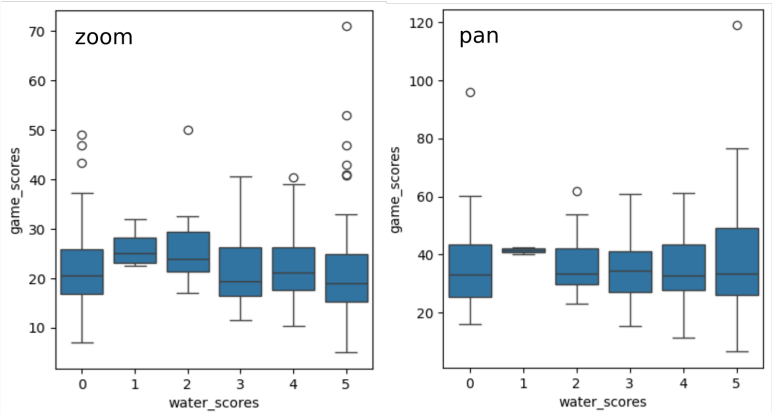


Figure 10 Box plot of the distribution of the number of zoom and pan interactions per game by water-level score.

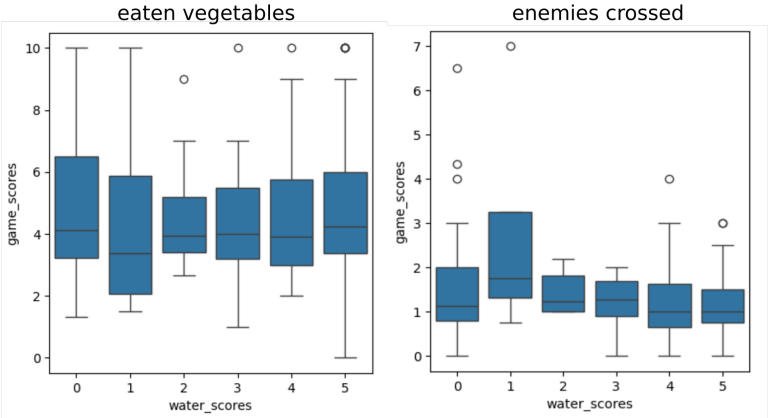


Figure 11 Box plot of the distribution of the number of bonuses and maluses obtained in the game (vegetables and enemies) by water-level score.

The game z-score is not the only metric that is collected during the completion of a game, as we record all the interactions of the player with the map. Figure 10 shows the box plots of the mean of zoom interactions per game by the water-level score, and of the mean of pan interactions per game by the water-level score. If we put aside the players who scored zero at the water-level task, we can see a weak correlation with the number of zoom interactions, confirmed by Spearman's correlation (-0.143; $p - value = 7.698 \cdot 10^{-2}$). Regarding the number of pans, Spearman's correlation is not significant (0.025; $p - value = 0.725$).

Figure 11 presents two additional box plots: (i) the number of vegetables consumed per game as a function of water-level category, and (ii) the number of enemies crossed per game as a function of water-level category. After excluding, once again, the players who obtained a score of zero on the water-level task, no statistically significant association is observed between water-level performance and the number of enemies crossed (Spearman's $\rho = -0.147$, $p = 6.890 \times 10^{-2}$), nor between water-level performance and the number of vegetables consumed (Spearman's $\rho = 0.061$, $p = 0.39$).

Finally, with respect to the number of games required to estimate players' skills with 95% confidence, the observed volatility for players who have participated in more than eight games corresponds to a z-score standard deviation of 0.71. We then computed the required

sample size for a margin of error slightly below this volatility and obtained that a minimum of three games is sufficient. When a smaller margin of error is imposed (0.5), the minimum number of games needed to achieve 95% confidence increases to seven.

4 Discussion

For now, the analysis of the data from the early players show a weak but significant correlation between the spatial skills measured with the water-level task, and the score in the game. But one reason why this correlation is weak is the lack of control in the study. Apparently, several players sidestepped the water-level task, maybe because it was boring compared to the game or because they did not understand it. The choice of a mobile game to study pan-scalar map-using skills was made to gather a large number of players, which is still our aim, but the drawback is the lack of control on how the users play the game and answer seriously to the spatial ability tests. From what we can see in the current results, the number of players sidestepping the water-level task is not very large (less than 5%), and we hope that with more players, the future analyses will be more robust.

With the score and the z-score, Mapinou measures the ability of players to be good at this game, which is not exactly similar to what we are primarily interested in: pan-scalar map-using skills. To perform at this game, players have to be able to search a target in a pan-scalar map, plan a route, and follow it, zoom in and out regularly, update your survey and route knowledge to collect vegetables. But being agile with your fingers and being able to very quickly chain sequences of finger interactions, is very useful to achieve a good score, like in many video games. It is questionable which type of skills are the most important to achieve good results in Mapinou, and maybe a future analysis of the speed of interaction will be necessary to address this question. There are studies in human-computer interaction about those touch-gestures on mobile devices that could be interesting to replicate with map zooming and panning tasks [31, 44, 45, 34].

Being crafty with touch-gestures on mobile devices can be somehow related to age. One of our initial hypotheses when designing Mapinou was that older users of pan-scalar maps might have more difficulties than young users. For now, this hypothesis has not been tested because the number of users older than 45 is too low to perform statistical analyses. However, our current results show that the "35-44" age category performs significantly better than the "25-34" category. This result shows that we should not expect a linear decline with age, but maybe we can expect a phenomenon of decay for older players. The results for the Sea Hero Quest mobile game show a peak of performance at 21 years old, and a global decrease as the age of the players advances [8]. Other studies show a non-linear effect of aging on spatial visualization and perspective taking cognitive competencies, and a linear effect on mental rotation [3].

Past studies on spatial abilities exhibit differences with gender [24, 6]. For instance, the results in the Sea Hero Quest mobile game show a significant gap between men and women, but the results also show that those differences highly depend on the country where the participants were born and raised, with countries such as Norway and Finland showing almost no gender difference in spatial ability, while other countries exhibit strong gender differences. According to these results, France, where all our participants live, show momderate gender differences, and this is what we can see in our results too. The distribution of z-scores are significantly different but the median of z-scores is almost identical for men and women. The difference comes from the higher number of very good scores among the male players, and the significant number of bad scores among female players. When we have more players to analyze, it would be interesting to see of this gender difference changes with age categories.

A previous study on Sea Hero Quest shows that the number of hours spent on video gaming predicts success in the game [51]. Playing video games has been shown to improve spatial skills [12, 10, 29], so we expect a similar effect on pan-scalar map using skills, and on the score in our game. We added a question in our demographic questionaire to indicate the number of hours per week spent playing video games on all devices, and we plan to study this effect in future analyses of the game data.

Socioeconomic status and educational attainment have also been shown to modulate spatial abilities [52]. Our initial recruitment procedure, which relied on posters displayed throughout the university campus, may therefore have yielded a relatively homogeneous sample in terms of education level. With a more educationally diverse sample, the proportion of participants exhibiting low water-level task performance might increase, potentially altering the pattern of results reported here. To mitigate this limitation in future data collection, we intend to diversify our recruitment strategies (e.g., via social media and public science outreach events) and to systematically collect detailed information on participants' socioeconomic status and educational background in our demographic questionnaire, in order to characterize and intentionally increase sample heterogeneity.

Finally, it has been shown that training to the use of maps has positive consequences on spatial cognition and spatial ability [47, 19]. Our results on the influence of self-declared map usage confirm this assertion, as players that use pan-scalar maps everyday have significantly better results than players who use them several times a week. We will try to collect as many players as possible with a use of pan-scalar maps that is less than several times a week to confirm this tendency with our other groups, even if it will not be easy because the game requires a mobile phone, and mapping applications are usually among the most used apps by mobile phone owners.

5 Conclusion

To conclude, this paper, based on the early players of Mapinou, our serious geo-game for mobile phones, shows that mobile games can be interesting tools to study the use of pan-scalar maps. Our first results show a weak but significant relation between the score at the water-level task (that is related to spatial ability) and the score in the game, which shows that players with better spatial abilities tend to perform better in the game.

To advance this work, we plan to recruit a substantially larger number of participants over the coming weeks in order to collect sufficient data for the three remaining spatial ability tests. Several communication actions are planned to favor this recruitment process. If the current findings generalize to these three additional tests, an open question is whether one of them exhibits a stronger correlation with game performance than the others, and thus could be preferentially adopted as a proxy measure of pan-scalar map abilities in subsequent user studies.

For one of the in-game spatial ability assessments, we incorporated a reduced subset (5 items) of the Purdue Spatial Visualization Test. Prior research indicates that scores on the full test are associated with participants' mental rotation abilities [15]. If the in-game scores show a substantial correlation with the scores on these 5 Purdue items, it will be necessary to verify that this reduced subset provides an adequate approximation of the full instrument. To this end, we plan to conduct a small-scale user study to evaluate the extent to which this abbreviated version serves as a valid proxy for the complete test.

Beyond the analysis of a dataset with a larger pool of players, we plan to reuse the game in future studies about pan-scalar cartography as it could be the first test that captures visual spatial ability and the ease to use pan and zoom on mobile devices. We also plan to study the Mapinou players with eye-tracking to better understand the cognitive and visual attention processes at play when zooming and panning in a pan-scalar map.

References

1 Vyron Antoniou and Christoph Schlieder. Participation Patterns, VGI and Gamification. In *Proceeding of: 17th AGILE Conference on Geographic Information Science - Geogames and Geoplay Workshop*, Castellon, Spain, June 2014. AGILE.

2 Leilani A. Arthurs, Sarah P. Baumann, Joel M. Rice, and Shelby Dianne Litton. The development of individuals' map-reading skill: what research and theory tell us. *International Journal of Cartography*, 9(1):3–28, January 2023. doi:10.1080/23729333.2021.1950318.

3 Erika Borella, Chiara Meneghetti, Lucia Ronconi, and Rossana De Beni. Spatial abilities across the adult life span. *Developmental Psychology*, 50(2):384–392, 2014. Place: US Publisher: American Psychological Association. doi:10.1037/a0033818.

4 James R. Carter. The Many Dimensions of Map Use. In *Proceedings of 22nd International Cartographic Conference*, pages 1–9, La Coruna, Spain, 2005. ICA.

5 Jésus Castellote, Joaquin Huerta, Javier Pescador, and Michael Brown. Town conquer : a gamified application to collect geographical names. In Joaquin Huerta, editor, *Proceedings of the AGILE'2014 International Conference on Geographic Information Science*, Castellón, Spain, June 2014. AGILE.

6 A. Coutrot, E. Manley, S. Goodroe, C. Gahnstrom, G. Filomena, D. Yesiltepe, R. C. Dalton, J. M. Wiener, C. Hölscher, M. Hornberger, and H. J. Spiers. Entropy of city street networks linked to future spatial navigation ability. *Nature*, 604(7904):104–110, April 2022. doi:10.1038/s41586-022-04486-7.

7 Antoine Coutrot, Sophie Schmidt, Lena Coutrot, Jessica Pittman, Lynn Hong, Jan M. Wiener, Christoph Hölscher, Ruth C. Dalton, Michael Hornberger, and Hugo J. Spiers. Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PLOS ONE*, 14(3):e0213272, March 2019. doi:10.1371/journal.pone.0213272.

8 Antoine Coutrot, Ricardo Silva, Ed Manley, Will de Cothi, Saber Sami, Véronique D. Bohbot, Jan M. Wiener, Christoph Hölscher, Ruth C. Dalton, Michael Hornberger, and Hugo J. Spiers. Global Determinants of Navigation Ability. *Current Biology*, 28(17):2861–2866.e4, September 2018. doi:10.1016/j.cub.2018.06.009.

9 Jeremy W. Crampton. Interactivity Types in Geographic Visualization. *Cartography and Geographic Information Science*, 29(2):85–98, January 2002. doi:10.1559/152304002782053314.

10 Jing Feng, Ian Spence, and Jay Pratt. Playing an Action Video Game Reduces Gender Differences in Spatial Cognition. *Psychological Science*, 18(10):850–855, October 2007. Publisher: SAGE Publications Inc. doi:10.1111/j.1467-9280.2007.01990.x.

11 Alinda Friedman, Bernd Kohler, Peri Gunalp, Alexander P. Boone, and Mary Hegarty. A computerized spatial orientation test. *Behavior Research Methods*, 52(2):799–812, April 2020. doi:10.3758/s13428-019-01277-3.

12 C. Shawn Green and Daphne Bavelier. Action video game modifies visual selective attention. *Nature*, 423(6939):534–537, May 2003. Publisher: Nature Publishing Group. doi:10.1038/nature01647.

13 Anita Greenhill, Kate Holmes, Jamie Woodcock, Chris Lintott, Brooke D Simmons, Gary Graham, Joe Cox, Eun Young Oh, and Karen Masters. Playing with science: Exploring how game activity motivates users participation on an online citizen science platform. *Aslib Journal of Information Management*, 68(3):306–325, 2016. doi:10.1108/AJIM-11-2015-0182.

14 Maïeul Gruget, Guillaume Touya, and Ian Muehlenhaus. Missing the city for buildings? A critical review of pan-scalar map generalization and design in contemporary zoomable maps. *International Journal of Cartography*, 9(2):255–285, January 2023. doi:10.1080/23729333.2022.2153467.

15 R. B. Guay. Purdue Spatial Visualization Test - Visualization of Rotations, 1977.

16 Mary Hegarty, Daniel R. Montello, Anthony E. Richardson, Toru Ishikawa, and Kristin Lovelace. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2):151–176, March 2006. doi:10.1016/j.intell.2005.09.005.

17 Mary Hegarty, Anthony E Richardson, Daniel R Montello, Kristin Lovelace, and Ilavanil Subbiah. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, September 2002. doi:10.1016/S0160-2896(02)00116-2.

18 Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, March 2004. doi:10.1016/j.intell.2003.12.001.

19 Toru Ishikawa. Individual Differences and Skill Training in Cognitive Mapping: How and Why People Differ. *Topics in Cognitive Science*, 15(1):163–186, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12605. doi:10.1111/tops.12605.

20 Susanne Jul and George W. Furnas. Critical Zones in Desert Fog: Aids to Multiscale Navigation. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST '98, pages 97–106, New York, NY, USA, 1998. ACM. event-place: San Francisco, California, USA. doi:10.1145/288392.288578.

21 Peter Kiefer, Ioannis Giannopoulos, and Martin Raubal. Using eye movements to recognize activities on cartographic maps. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 488–491, New York, NY, USA, November 2013. Association for Computing Machinery. doi:10.1145/2525314.2525467.

22 Joel Kovarsky. The Many Uses of Maps: The Producer at the Center. *Cartographic Perspectives*, 57:12–20, 2007. doi:10.14714/CP57.278.

23 Maria Kozhevnikov and Mary Hegarty. A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29(5):745–756, July 2001. doi:10.3758/BF03200477.

24 Lynn S. Liben and Susan L. Golbeck. Sex Differences in Performance on Piagetian Spatial Tasks: Differences in Competence or Performance? *Child Development*, 51(2):594–597, 1980. Publisher: [Wiley, Society for Research in Child Development]. doi:10.2307/1129301.

25 Lynn S. Liben, Lauren J. Myers, and Kim A. Kastens. Locating Oneself on a Map in Relation to Person Qualities and Map Characteristics. In Christian Freksa, Nora S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, pages 171–187, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. doi:10.1007/978-3-540-87601-4_14.

26 Amy K. Lobben. Tasks, Strategies, and Cognitive Processes Associated With Navigational Map Reading: A Review Perspective. *The Professional Geographer*, 56(2):270–281, May 2004. doi:10.1111/j.0033-0124.2004.05602010.x.

27 Ian McCallum, Linda See, Tobias Sturn, Carl Salk, Christoph Perger, Martina Duerauer, Mathias Karner, Inian Moorthy, Dahlia Domian, Dmitry Schepaschenko, and Steffen Fritz. Engaging Citizens in Environmental Monitoring via Gaming. *International Journal of Spatial Data Infrastructures Research*, 13:15–23, 2018. URL: http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/474/416.

28 Judith M. W. Meyer. Map Skills Instruction and the Child's Developing Cognitive Abilities. *Journal of Geography*, 72(6):27–35, September 1973. doi:10.1080/00221347308981316.

29 Kara Murias, Kathy Kwok, Adrian Gil Castillejo, Irene Liu, and Giuseppe Iaria. The effects of video game use on performance in a virtual navigation task. *Computers in Human Behavior*, 58:398–406, May 2016. doi:10.1016/j.chb.2016.01.020.

30 Carman Neustaedter, Anthony Tang, and TejinderK Judge. Creating scalable location-based games: lessons from Geocaching. *Personal and Ubiquitous Computing*, 17(2):335–349, February 2013. doi:10.1007/s00779-011-0497-7.

31 Alexander Ng, John Williamson, and Stephen Brewster. The Effects of Encumbrance and Mobility on Touch-Based Gesture Interactions for Mobile Phones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, pages 536–546, New York, NY, USA, August 2015. Association for Computing Machinery. doi:10.1145/2785830.2785853.

32 Jean Piaget and Bärbel Inhelder. *The Child's Conception of Space*. Routledge & Kegan Paul, London, 1956.

33 Tong Qin, Wim Fias, Nico Van de Weghe, and Haosheng Huang. Differentiating Everyday Map Tasks: Unique Attention-Related Eye Movements and Electrophysiological Signatures of Map Use. *Annals of the American Association of Geographers*, 114(7):1604–1624, 2024. doi:10.1080/24694452.2024.2353845.

34 Yosra Rekik, Radu-Daniel Vatavu, and Laurent Grisoni. Understanding users' perceived difficulty of multi-touch gesture articulation. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 232–239, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2663204.2663273.

35 Robert E. Roth. An Empirically-Derived Taxonomy of Interaction Primitives for Interactive Cartography and Geovisualization. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2356–2365, 2013. doi:10.1109/TVCG.2013.130.

36 Ethan Russell. 9 things to know about Google's maps data: Beyond the Map, 2019. URL: https://mapsplatform.google.com/resources/blog/9-things-know-about-googles-maps-data-beyond-map/.

37 C. Schlieder, P. Kiefer, and S. Matyas. Geogames: Designing Location-Based Games from Classic Board Games. *Intelligent Systems, IEEE*, 21(5):40–46, September 2006. doi:10.1109/mis.2006.93.

38 Angela Schwering, Jakub Krukar, Jana Seep, and Yousef Qamaz. Individualization in Spatial Behaviour and Map Reading. *AGILE: GIScience Series*, 4:1–5, June 2023. doi:10.5194/agile-giss-4-41-2023.

39 Linda See, Tobias Sturn, Christoph Perger, Steffen Fritz, Ian McCallum, and Carl Salk. Cropland Capture: A Gaming Approach to Improve Global Land Cover. In Joaquin Huerta, Sven Schade, and Carlos Granell, editors, *Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science*, Castellón, Spain, June 2014. AGILE.

40 Hugo J Spiers, Antoine Coutrot, and Michael Hornberger. Explaining World-Wide Variation in Navigation Ability from Millions of People: Citizen Science Project Sea Hero Quest. *Topics in cognitive science*, 15:120–138, 2023. Publisher: Wiley. doi:10.1111/tops.12590.

41 Csaba Szigeti-Pap, Dávid Kis, and Gáspár Albert. Understanding Map Misinterpretation: Factors Influencing Correct Map Reading and Common Errors. *ISPRS International Journal of Geo-Information*, 12(12), November 2023. doi:10.3390/ijgi12120479.

42 Guillaume Touya and Justin Berli. The FogDetector: A User Survey to Measure Disorientation in Pan-Scalar Maps. In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 73:1–73:6, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.GIScience.2023.73.

43 Guillaume Touya, Maïeul Gruget, and Ian Muehlenhaus. Where Am I Now? Modelling Disorientation in Pan-Scalar Maps. *ISPRS International Journal of Geo-Information*, 12(2):62, February 2023. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/ijgi12020062.

44 Jessica J. Tran, Shari Trewin, Calvin Swart, Bonnie E. John, and John C. Thomas. Exploring pinch and spread gestures on mobile devices. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '13, pages 151–160, New York, NY, USA, August 2013. Association for Computing Machinery. doi:10.1145/2493190.2493221.

45 Shari Trewin, Cal Swart, and Donna Pettick. Physical accessibility of touchscreen smartphones. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 1–8, New York, NY, USA, October 2013. Association for Computing Machinery. doi:10.1145/2513383.2513446.

46 Jean D. M. Underwood. SkiIled map interpretation and visual-spatial ability. *Journal of Geography*, 80(6):55–58, February 1981. doi:10.1080/00221348108980235.

47 David H. Uttal. Seeing the big picture: map use and the development of spatial cognition. *Developmental Science*, 3(3):247–264, 2000. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-7687.00119. doi:10.1111/1467-7687.00119.

48 Steven G. Vandenberg and Allan R. Kuse. Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and Motor Skills*, 47(2):599–604, December 1978. doi:10.2466/pms.1978.47.2.599.

49 Michele Andrisin Wittig and Mary J. Allen. Measurement of adult performance on Piaget's water horizontality task. *Intelligence*, 8(4):305–313, October 1984. doi:10.1016/0160-2896(84)90014-X.

50 Thomas Wolbers and Mary Hegarty. What determines our navigational abilities? *Trends in Cognitive Sciences*, 14(3):138–146, March 2010. Publisher: Elsevier. doi:10.1016/j.tics.2010.01.001.

51 Emre Yavuz, Chuanxiuyue He, Sarah Goodroe, Chris Ganstrom, Antoine Coutrot, Michael Hornberger, Mary Hegarty, and Hugo J. Spiers. Video gaming, but not reliance on GPS, is associated with spatial navigation performance. *Journal of Environmental Psychology*, 96:102296, June 2024. doi:10.1016/j.jenvp.2024.102296.

52 Emre Yavuz and Hugo J. Spiers. Determinants of individual navigation ability. *Nature Reviews Psychology*, 5(4):273–289, April 2026. Publisher: Nature Publishing Group. doi:10.1038/s44159-026-00544-3.

An Operator Framework for Real-Time Analytics of Streamed Fields

JJ Paul
Spatial Computing, SCIS, University of Maine, Orono, ME, USA

Silvia Nittel
Spatial Computing, SCIS, University of Maine, Orono, ME, USA

Abstract

Continuous geospatial phenomena such as air quality, temperature, soil moisture, or noise evolve smoothly over space while potentially changing rapidly over time. They are increasingly observed through dense, real-time measurement streams from geosensor networks, mobile platforms, and live sensor infrastructures. While data stream engines provide low-latency processing over unbounded streams, they lack higher-level abstractions for spatially continuous phenomena. As a result, real-time field analysis is often implemented via ad-hoc pipelines that combine GIS tools, stream processing, and application-specific logic. In this paper we propose an analytical operator framework for streamed scalar fields. We derive a minimal, composable operator set building on principles of compact operator sets and dimensional reasoning over space, time, and theme. We demonstrate how analytics for streamed continuous phenomena are expressed through primitive operator compositions. The resulting framework provides a systematic basis for describing and implementing extensible stream-based, real-time field analytics.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Geostreaming, Spatiotemporal phenomena, Fields, Data streams, Real-time analytics, Operators, Algebra

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.14

1 Introduction

Geospatial phenomena evolve continuously across space and time [18, 37]. Air quality, temperature, soil moisture, and noise levels vary smoothly across spatial domains while undergoing rapid temporal change [19, 34]. These phenomena are increasingly observed through dense streams from wireless geosensor networks, mobile sensing platforms, and live data infrastructures [29, 28, 1]. As sensing technologies proliferate, spatial and temporal coverage increases, yielding large volumes of continuously arriving measurements [23, 14].

Despite improved observational capacity, two core challenges remain. First, continuous phenomena require high-level data representations beyond raw atomic measurements [25, 8]. Second, real-time settings demand analytical frameworks that support automated decision making and operational response [33]. Analysis must proceed incrementally as phenomena unfold [23]. These requirements differ from traditional retrospective Geographic Information Science (GIScience) analysis and from conventional data stream analytics [4, 44]. In particular, analytical processes must operate progressively under strict latency requirements, continuously updating results as new measurements arrive while respecting the spatial continuity and thematic structure characteristic of phenomena [6].

Environmental monitoring illustrates this challenge. In precision agriculture, soil moisture and nutrient levels are measured in real-time as machinery traverses cropland, requiring immediate synthesis to guide irrigation and fertilization [45]. City-wide sensor networks stream air quality measurements that must be continuously aggregated and transformed to detect localized hazards [31, 32]. For example, city-wide air-quality streams may be

© JJ Paul and Silvia Nittel;
licensed under Creative Commons License CC-BY 4.0

filtered or joined to districts as sensor events. In such settings, analysis proceeds forward in time and must reflect the current state of the phenomenon without complete historical access [4, 23, 44, 2].

Existing analytical frameworks for spatiotemporal phenomena [15, 1] inadequately support real-time field-based analysis. GIS provides a set of extensive operators for spatial fields but they assume static or periodically updated datasets and full access to historical data [15, 37, 26, 8]. Data stream engines (DSEs) support high-throughput, low-latency processing over unbounded streams yet generally operate on flat relational abstractions lacking higher-level representations for spatially continuous fields [4, 6, 44]. Prior geostreaming research has largely focused on moving objects as in PLACE [27], StreamInsight [21], and Galić et al. [16], while work on continuous phenomena has been more limited [42, 25, 2]. Consequently, real-time analysis over continuous geospatial fields is commonly implemented through ad-hoc combinations of GIS tools, stream processing systems, and application-specific logic [24, 1, 23]. These solutions are technically demanding and difficult to generalize [1].

Existing approaches for detailed spatiotemporal continuous field analysis are based on time series analysis; this assumes a fixed time sequence of measurements, which can be revisited arbitrarily, enabling global statistics and retrospective segmentation [15, 37]. In contrast, real-time stream processing operates under an advancing notion of time in which future observations are unavailable and past observations may not be retained [5, 10]. Analysis must therefore be incremental and yield results without access to a complete historical copy of the input stream [23, 44]. At present, there is no broadly accepted compact operator framework that unifies continuous spatial fields, spatiotemporal reasoning, and stream-based constraints into a single analytical vocabulary.

Contributions. This paper introduces a novel operator set for the analysis of streamed-based, spatiotemporal scalar fields. Its design builds on dimensional reasoning over space, time, and theme and integrates the computational operating principles of stream processing. The framework provides a compact, minimal, and expressive set of operators enabling a systematic basis for describing and implementing extensible real-time analytics over spatiotemporal continuous phenomena sampled by spatially and temporally dense sensing environments.

This article is organized as follows: we discuss the background of continuous spatiotemporal phenomena analysis in Section 2. Section 3 introduces our operator framework design, the Streamed Measurement Field (SMF) and the proposed analysis operator set. Section 4 formalizes the operators. Section 5 includes validation and discusses some limitations. Section 6 compares our operator framework with related approaches and we conclude in Section 7.

2 Background

2.1 Analysis of Continuous Spatiotemporal Phenomena

This paper focuses on the analysis of continuous, spatiotemporal phenomena. Fields have long been considered conceptual models of continuous phenomena; they represent functions from spatiotemporal locations to qualities rather than depicting implementation formats [17, 22]. The concept of fields has mostly been used to describe continuous phenomena. Coverages are conceptually similar to fields but have primarily been used to describe data structures for continuous phenomena, providing a more generic data type than rasters or tessellations [13, 19]. In this article, however, we also see fields as data types [8, 24, 35]. Furthermore, we focus specifically on scalar fields, which are functions assigning a single value to each location in

space and time, representing the intensity of a phenomenon without directional components. In practice, such fields are observed through discrete measurements at sensor locations at points in time, sampling phenomena such as temperature, air quality, or soil moisture [37, 19]. Here, continuous refers to the conceptual view of a phenomenon as varying continuously over space and time, while dense refers to frequent and proximate samples that support near-real-time field reasoning [42, 20]. Further, we distinguish between observed and derived fields. An observed field consists of discrete, spatiotemporal measurement samples, and a derived field results from interpolation or estimation approximating values over unsampled locations to generate e.g. rasters. Thus, scalar fields and their variation across space and time are the analytical object of interest in this paper.

Historically, analytical capabilities over continuous phenomena were constrained by sparse observations, often in space and time. Early analyses operated on fields observed at discrete time steps, emphasizing spatial interpolation, smoothing, and spatial pattern characterization at individual instants. Change was typically computed by comparing successive static states, following a "Time-Slice Snapshot" model of temporal reasoning [37, 15]. As sensing technologies improved, denser temporal sampling enabled more advanced temporal analysis. Much of this work shifted toward representations centered on moving objects and their trajectories [1]. Although such models support reasoning about duration and ordering, they emphasize persistent identities and discrete trajectories rather than continuous variation over space; the analytical focus remains object-centric.

In recent years, technological advances have produced denser observations across space and time, yielding spatiotemporal scalar fields [29]. Now, analysis includes joint reasoning over space and time: examining evolving spatial patterns, neighborhood change, and region-dependent trends. Rather than isolated snapshots or object-centric histories, analysis operates over continuously varying fields, supporting more faithful representations of dynamic conditions [1]. With geosensor networks and IoT-equipped sensing, sensor streams mark a further advancement. Observations arrive continuously, at high volume and in real-time, enabling analysis as phenomena unfold rather than retrospectively [23, 30]. Dense, real-time observations invite questions about short-term change, local dynamics, and evolving spatial structure that were previously impractical. These observations enable real-time decision making, as found in urban air quality monitoring. However, they require rethinking field-based analysis under monotonic time progression.

2.2 Stream-based Data Acquisition for Spatiotemporal Phenomena

Stream-based analytics takes place as phenomena are observed in real-time [2, 42]. Observations arrive incrementally with minimal delay as sequences of measurements and analyses are immediately and continuously re-evaluated as new data become available. Results are produced relative to t_{now}, rather than over a fixed temporal interval of the past as assumed in traditional spatiotemporal analysis. A defining characteristic of real-time stream-based analytics is explicit management of analytical state [5]. Measurement streams are unbounded, meaning we cannot predict how much data will arrive in the future. At any moment, only a finite subset of data has arrived. DSEs support forward-looking analyses and retain very little historical information [6, 5, 10]. Temporal context is specified through windows defined relative to t_{now} [5]. More general stream algebras define additional window types [36]. In this article, we focus on time-based windows because they match the advancing-time semantics of sensor streams and remain compatible with common data stream engines.

2.3 Analysis via Data Dimensions

A foundational principle in GIS is that analytical intent can be described through dimensions of space, time, and theme [39]. Rather than enumerating specific analysis tasks directly, this perspective characterizes analysis by which dimensions are fixed, controlled, or measured [39]. For example, fixing time while varying location and measuring a quantity such as temperature yields a spatial field representing continuous temperature over a region at a specific moment. Such a representation enables reasoning about spatial variation under a fixed temporal context.

When applied to scalar fields, dimensional reasoning concisely expresses analytical intent without prescribing algorithms and is therefore declarative. A temperature field may be restricted to a spatial subregion or summarized over a recent interval. Measuring spatial variation at a fixed time, summarizing values over time at fixed locations, or examining evolving spatial gradients are all instances of different dimensional configurations of the same underlying spatiotemporal field [39, 15]. This abstraction foregrounds the type of reasoning being performed rather than its implementation. When analyzing a single phenomenon, theme is fixed, but metric observations can be varied or measured.

Similar dimension-oriented thinking also underlies applied systems such as On-Line Analytical Processing (OLAP) [12]. OLAP organizes analysis into multidimensional data tuples (e.g., sales records). Although OLAP does not natively support continuous spatial data, its success illustrates the expressive power of dimensional organization for summarizing and exploring datasets consisting of multidimensional tuples [12].

This view remains meaningful in stream-based settings. The dimensions of space, time, and theme do not disappear when data arrive as unbounded streams [23]. The challenge lies in determining which forms of dimensional reasoning remain well-defined under monotonically advancing time [5, 12]. Grounding analysis in data dimensions rather than specific operations provides a flexible foundation for reasoning about continuous phenomena across observational regimes.

2.4 Analysis Operator Frameworks for Spatiotemporal Fields

Across analytical traditions, a central design principle is that complex behavior can be expressed through the composition of a compact set of primitive operations [11, 41]. Rather than treating each task as a special case, analysis is organized through operator sets that function as a general analytical vocabulary. This principle has proven effective across domains and motivates its extension to continuous spatiotemporal phenomena.

For instance, the relational algebra [11] demonstrated that a limited collection of set-based operators is sufficient to express diverse analytical queries through composition and boolean logic. The emphasis on composability and expressive sufficiency shaped analytical systems such as SQL and other relational algebra or relational calculus languages [9, 4], providing a foundation for high-level data languages.

For continuous spatial fields, Map Algebra provides a comparable vocabulary through local, focal, and zonal operations, enabling the derivation of new fields, neighborhood analysis, and regional aggregation [41]. Extensions such as Cubic Map Algebra generalized these ideas to incorporate time, enabling operations over three-dimensional space-time grids. However, such frameworks largely assume fully materialized rasters or regularly sampled grids and remain primarily spatial in orientation. While influential in GIS practice, they do not constitute a widely adopted, unified operator framework for continuous spatiotemporal fields and they do not support stream processing.

Subsequent work has explored spatiotemporal algebras and operator sets that integrate spatial and temporal reasoning [15, 25]. These approaches facilitate subsetting and statistical reduction over multidimensional grids and emphasize temporal slicing, aggregation, and change analysis. Time series analysis shares this operator logic in that it supports filtering, segmentation, and statistical summarization over ordered observations. However, both time series and many spatiotemporal algebraic frameworks presuppose complete historical access to the temporal extent.

2.5 Limitations of Existing Approaches

Stream-based analytics differs in its computational model. Time series analysis typically assumes a fixed sequence that can be revisited arbitrarily, enabling global statistics and retrospective segmentation [15]. In contrast stream processing operates under an advancing notion of time in which future observations are unavailable and past observations may not be retained [5]. Analysis must therefore be incremental and yield results without access to a complete copy of the input stream [23]. Current geostreaming focuses on moving objects, spatial tuples, and vector geometries [27, 16, 44].

At present, there is no broadly accepted compact operator framework that unifies continuous spatiotemporal fields, spatiotemporal reasoning, and stream-based constraints into a single analytical vocabulary. Existing operator sets depend on representational assumptions and are largely retrospective [15]. The integration of spatial continuity with forward-only temporal semantics remains an open challenge. The combination of spatial and temporal operators under streaming constraints is therefore the central issue.

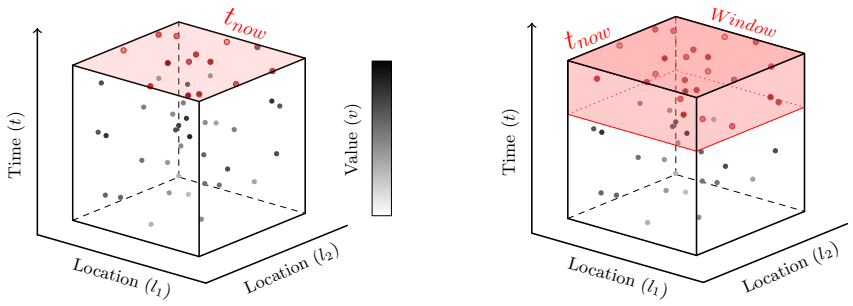
3 An Analytical Operator Framework for Streamed Spatiotemporal Scalar Fields

This section introduces our proposed analytical operator framework for streamed spatiotemporal scalar fields.

3.1 Rationale

In this paper, we call representations of continuous ST phenomena that are sampled by measurement streams a streamed scalar field (S-SF). Our design was driven by the following considerations:

- First, we focused on identifying the *analytical tasks* for streamed scalar fields. An analytical task describes a class of reasoning performed over continuous spatiotemporal phenomenon. Adopting a task-oriented view provides a stable abstraction for assessing analytical expressiveness and allows us to be compatible with other forms of analyzing spatiotemporal fields such as fields based on historic, time-series data.
- The second design principle identifies a *minimal set of primitive operators* [11, 41] to support the analytical tasks of Step 1 rather than building an exhaustive collection of specialized operations. We expect this operator set to be sufficient for expressing complex pattern analytics for continuous phenomena through composition while adhering to stream-based, real-time semantics.
- The third principle targets the *data dimensions* of S-SF (spatial, temporal, and thematic aspects). Our intent is to provide operator primitives that can be equally applied along the different data dimensions of fields.



(a) SMF at t_{now}.

(b) Window over Streamed Measurement Field.

Figure 1 Streamed Measurement Field Representation.

3.2 The Streamed Measurement Field Data Model

When analyzing continuous phenomena that are captured by streams of measurements, incoming observations are treated as samples of an underlying phenomenon rather than as independent events, which allows analysis to preserve spatial continuity while remaining compatible with real-time stream processing. To represent such streamed fields, we adopt the Streamed Measurement Field Model (SMF) [35]. The SMF data model builds on atomic observations [19] and systematically defines increasingly structured data types: measurements, measurement streams, measurement fields, and streamed measurement fields. We present a selected subset of data types as basis for the operator framework.

▶ **Definition 1** (Measurement). *A measurement $m = ((l,t), Z, z(l,t))$ represents the value of a property Z at a spacetime point $x = (l,t) \in D = L \times T$. The function $z(l,t)$ maps the spacetime point (l,t) to a value of Z. The metric observation $z(l,t)$ is some value $v \in \mathbb{R}$.*

For simplicity, the formal model uses real-valued spatial, temporal, and measurement domains. Implementations necessarily approximate these domains through finite-precision computational representations [38]. We use measurement subscripts to denote component projection, so l_m, t_m, and Z_m denote the location, time, and property of measurement m. Numeric subscripts denote ordered instants, as in $t_0, t_1, \ldots$, while F_t denotes the field induced at time t. A scalar field is classically defined as a function $f : L \times T \to \mathbb{R}$, mapping each spacetime coordinate to a value [19]. A measurement field instantiates this concept over a finite measurement set.

▶ **Definition 2** (Measurement Field). *A measurement field is defined by a set of measurements M, a property Z, and a spacetime domain D.*

$$F = (M_F, Z_F, D_F) \quad \textit{where} \quad \forall m \in M_F \mid Z_m = Z_F \wedge l_m \in L_F \wedge t_m \in T_F.$$

Sensor-generated measurements arrive as a time-ordered stream $S \to \langle m_{t_0}, m_{t_1}, \ldots, m_{t_{\text{now}}} \rangle$ [30, 4]. The SMF constrains the stream to a single property Z and spatial domain L. This represents evolving observations of a given phenomenon over time.

▶ **Definition 3** (Streamed Measurement Field). *A streamed measurement field is defined by a measurement stream S, property Z, and location domain L which together indirectly induce a snapshot measurement field F at each instant t_{now}.*

$$SMF = (S, Z, L) \to \langle F_{t_0}, F_{t_1}, \ldots, F_{t_{now}} \rangle$$

where each snapshot measurement field is

$$F_{t_{now}} = \{m \in S \mid Z_m = Z_F \wedge l_m \in L_F \wedge t_m = t_{now}\}.$$

Stream processing requires a bounded state. This is achieved through windowing which defines finite temporal evaluation contexts relative to the present t_{now} [6, 4]. A temporal window selects measurements whose timestamps fall within a specified interval, yielding a finite measurement field F suitable for analysis.

▶ **Definition 4** (Window). *Let $W(i,k)$ denote a window specification with interval length $i > 0$ and slide $k > 0$. The window operator is defined as*

$$\mathsf{Window}_{(i,k)} : SMF \to \langle F_{w,0}, F_{w,1}, F_{w,2}, \dots \rangle,$$

where each window is indexed by an integer $j \in \mathbb{N}$, and the induced measurement field is

$$F_{w,j} = \{\, m \in S \mid t_m \in I_j \,\} \quad \textit{where} \quad I_j = [t_0 + jk, t_0 + jk + i).$$

Temporal window configurations include tumbling (non-overlapping), sliding (overlapping), and hopping (gapped), and provide a mechanism for temporal reasoning.

3.3 Defining a Core Analysis Task Set

3.3.1 Rationale

The construction of a core analysis task set begins by synthesizing insights from several established analytical traditions that address different aspects of spatiotemporal reasoning [39, 37, 11, 4, 23]. These traditions include field-based analysis in GIS, dimensional perspectives on geographic information, spatiotemporal data models, relational algebra, and stream processing semantics [39, 19, 15, 6]. Established mechanisms appear under different names such as map, combine, and reduce, but they reflect similar forms of analytical intent rooted in the fundamental relationships between what, where, and when [37, 8, 19]. The synthesis makes explicit that the underlying assumptions of these frameworks are not uniform [37, 4, 15]. Retrospective GIS and spatiotemporal models typically assume access to unrestricted history, with time series as prevalent model [37, 15]. Stream processing models instead assume unbounded data and forward-only evaluation [4, 23].

1. One common behavior is the *restriction* of analytical scope. Analysis is often limited to a subset of space, a bounded temporal interval, or a subset of values satisfying a specific condition [39, 11]. This behavior focuses attention on a relevant portion of the phenomenon, without altering the underlying structure of the observations [26, 23].
2. A second recurring behavior is the organization of data into *finite units* of analysis. Continuous phenomena are frequently partitioned into spatial regions, temporal intervals, or value-based categories to make analysis tractable [19, 37].
3. A third behavior is the *summarization* of values within units of analysis. Aggregation operations compute statistics, indicators, or descriptive measures that characterize the phenomenon within each unit [1, 12]. This allows reasoning about local variation while reducing data volume through operations like zonal statistics or global means [41, 26].
4. A fourth recurring behavior is the *transformation* of measurements or fields to derive new representations. Transformations may map values to different units, normalize measurements, or apply functional mappings [8, 41]. Spatially and temporally, transformation allows representation under different projections or encodings.
5. A fifth behavior is the *correlation* or joining of multiple phenomena to examine their interactions. This behavior enables joint reasoning over distinct fields [4, 46].

6. A sixth recurring behavior is the recognition of *patterns*, events, or complex trajectories from the underlying field. This involves extracting discrete episodes with definite beginnings and ends.
7. A seventh recurring behavior is *numeric* derivation which employs basic mathematical primitives such as sum, difference, product, and ratio to compute a new value per unit of observation. These numeric operators are essential for transforming representation of a phenomena or establishing a mathematical relationship between two or more phenomena [41, 8, 19].
8. A final behavior is *neighborhood reasoning*, where the value at a given location is computed based on the context of its spatial or temporal surroundings [41, 26]. These focal operations characterize neighborhood structure, calculate gradients, or estimate distance decay effects within the field [41, 19].

Behaviors that require unrestricted access to the temporal extent of the data are excluded [15, 23]. Analyses that depend on future observations, global temporal segmentation, or the repeated traversal of historical data cannot be evaluated incrementally in stream processing [5, 23, 6, 23]. Furthermore, blocking operators such as sorting or global aggregation are inadmissible because they are unable to produce any output until an infinite stream has reached its end [6, 4, 7].

3.3.2 The Core Task Set

Applying unifying analytics of S-SF along data dimensions with streams yields a set of analytical operator classes that recur across spatial, temporal, relational, and streaming traditions and remain well-defined under forward-only evaluation and bounded state [4, 23, 46]. Each operator class represents a distinct mode of analytical intent that cannot be removed without reducing expressive capacity under stream semantics. We distinguish two families of operator classes (see Table 1):

Representation-Level Operators

Representation-Level Operator classes modify how a phenomenon is expressed. The ***Interpolate*** operator class estimates values at unsampled locations or times to produce a more continuous representation of the underlying scalar field. While interpolation may enable subsequent analysis, it primarily refines representation rather than collapsing dimensions or extracting summary descriptors. The Interpolate operator class has a function specification such as Kriging or Inverse Distance Weighting. The ***Transform*** class maps measurements to alternative spatial, temporal, or thematic encodings while preserving stream structure. Examples include coordinate projection, unit conversion, normalization, and temporal re-encoding.

Analysis-Level Operators

Analysis-Level Operator classes provide the primary means of deriving new insights from the SMF. Together, these seven task classes provide a minimal and composable vocabulary for reasoning over continuous spatiotemporal scalar fields in real-time environments [15, 8].

- **Decomposition** organizes measurements into finite units of analysis to ensure tractable computation over unbounded streams. Spatially, this corresponds to tessellations or regions; temporally, to windows; and metrically to binning or classification.

- **Restriction** limits analytical scope by selecting subsets of measurements according to spatial extent, temporal context, or value predicates. This preserves the underlying structure of measurements while focusing attention on relevant portions of a phenomenon.
- **Summarization** condenses measurements within units to derive statistics, indicators, or condensed representations. Expressed as zonal operations in map algebra or windowed reductions in stream engines, aggregation supports reasoning about trends and short-term variation [41, 4].
- **Numeric** tasks use primitive arithmetic operators on the values of measurements. These operators can be used to make adjustments, compare phenomena, or produce more complex analysis [41, 26].
- **Combination** supports joint reasoning across distinct streamed measurement fields of the same type. This includes spatial overlays and time-aware joins operating over synchronized windows to maintain semantic validity.

Table 1 Operator classes organized by analytical dimension and semantic level.

Operator	**Space** (L)	**Time** (T)	**Value/Theme** (V)
Representation-Level Operators			
Transform	Projection / Translation	Time encoding	Unit conversion
Interpolate	Fill missing locations	Fill missing instants	Fill missing values
Analysis-Level Operators			
Decompose	Spatial units	Window relative to T_{now}	Value classes
Restrict	Subset by spatial region	Measurements at T_{now}	Filter by value predicate
Numeric	Spatially identity	Temporal identity	Mathematical primitives
Summarize	Convex hull	Utilize windows	Metric statistics
Combine	Spatial membership	Synchronized instants	Metric membership

4 Operators

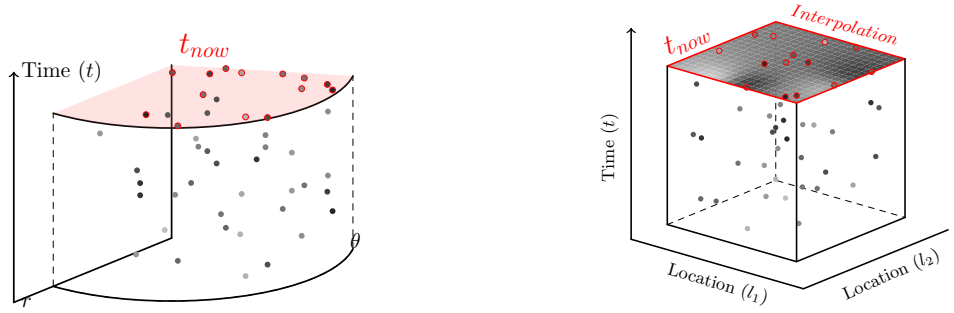
In this section we define the operator classes and their signatures, and examine whether and how they specialize by dimension. Figure 1a illustrates the base SMF upon which the operators act. Overall, we distinguish analytical operators, which extract new insight, and representation-level operators, which modify how the phenomenon is encoded or presented.

4.1 Representation-Level Operators

Representation-level operators often work directly on the measurement streams that constitute a SMF, for instance transform. Interpolation is also included since it produces continuous representations based on the raw measurements.

4.1.1 Transformation

Transformation (Figure 2a) is a measurement-level operator mapping one SMF to another while preserving its dimensional structure. It modifies the representation of space, time, or value without altering stream semantics. Let $\Delta : (T \times L \times V) \to (T' \times L' \times V')$ be a measurement-wise transformation. The operator has type: $\mathsf{Transform}_\Delta : SMF \to SMF'$.



(a) Spatial Transformation.

(b) Interpolation of SMF.

Figure 2 Examples of Representation-Level Operators.

4.1.1.1 Transformation by Dimension

Transformation modifies representation along a chosen dimension while preserving stream structure. Spatial transformation re-expresses locations in an alternative coordinate system (e.g., map projection). Temporal transformation changes how timestamps are encoded or displayed (e.g., calendar time to UNIX time) without altering stream progression relative to t_{now}. Thematic transformation re-expresses measurement values through unit conversion, normalization, or reclassification (e.g., Fahrenheit to Celsius). In all cases, the underlying spatiotemporal structure of the SMF is preserved; only the representation of individual SMF measurements are modified.

4.1.1.2 Example of Transformation (Unit Conversion)

For temperature, define $\Delta_V(v) = \frac{5}{9}(v-32)$. Applying $\mathsf{Transform}_{\Delta_V}$ converts all measurements from Fahrenheit to Celsius without altering spatial or temporal dimensions.

4.1.2 Interpolation

Interpolation (Figure 2b) is included because real-time field analysis requires estimation at unsampled locations or times. Steenbergen et al. [40] frame interpolation as deriving fields from measures, and Whittier et al. [42] show how such fields can be computed on-the-fly from real-time sensor observations. We therefore treat interpolation as a representation-level operator class that estimates values at target spacetime coordinates to produce a discretized representation of an underlying continuous scalar field.

Let F_w be a finite measurement field induced by a bounded window, and let $Q \subseteq L \times T$ be a set of target spacetime coordinates. An interpolation method κ, such as Inverse Distance Weighting (IDW) or Kriging, has type $I_\kappa : F_w \times Q \to F'_w$.

The output F'_w contains derived measurements at the target coordinates in Q, while preserving the property and domain of the input field. Spatial interpolation estimates values at unsampled locations, temporal interpolation estimates values at intermediate time instants, and spatiotemporal interpolation estimates values over target coordinates in $L \times T$. We do not define value-space interpolation as a separate scalar-field operator because scalar values are measures rather than controls.

4.1.2.1 Example of Spatial Interpolation (IDW)

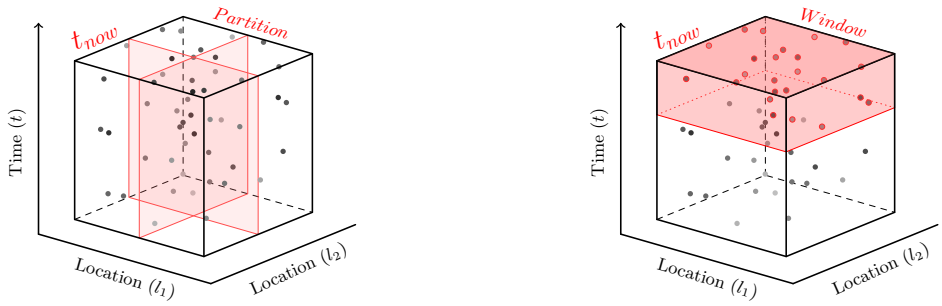
Let $v(l_i, t)$ be observed measurements. For an unsampled location l^*:

$$\hat{v}(l^*, t) = \frac{\sum_i w_i v(l_i, t)}{\sum_i w_i}, \quad w_i = \frac{1}{d(l^*, l_i)^p}.$$

Applying $\mathsf{Interpolate}_{\text{IDW}}$ estimates values at unsampled spatial locations while preserving temporal and thematic dimensions.

4.2 Analysis-Level Operators

4.2.1 Decomposition



(a) Spatial Partitioning of SMF.

(b) (Temporal) Window on SMF.

Figure 3 Decomposition Operators on SMF.

Decomposition organizes an SMF into evaluation units along a chosen data dimension. Let $D \in \{L, V, T\}$ denote a dimension (space, theme, or time), and let X_D be a decomposition specification over that dimension. Then: $\mathsf{Decompose}_D : SMF \times X_D \to \{F\}$, where $\{F\}$ is a collection of derived fields whose structure depends on D. Decomposition along spatial or metric dimensions produces streamed sub-SMFs. Decomposition along the temporal dimension produces finite Measurement Fields.

4.2.1.1 Spatial and Thematic Decomposition

When $D \in \{L, V\}$, decomposition divides the SMF into sub-SMFs according to a spatial tessellation (Figure 3a) or metric classification. Each sub-SMF remains an unbounded SMF evolving relative to t_{now}.

4.2.1.2 Temporal Decomposition (Windowing)

When $D = T$, decomposition is realized through windowing (Figure 3b). A window specification W defines bounded temporal intervals relative to t_{now}.

$$\mathsf{Decompose}_T \equiv \mathsf{Window} : SMF \times W \to \{F\}$$

The resulting F is a set of finite Measurement Fields. Each contains measurements whose timestamps fall within the window interval. If the decomposition specification induces pairwise disjoint and complete coverage of the parent SMF, the result satisfies classical partition properties. Windowing does not necessarily satisfy partition properties as hopping windows may leave gaps, and sliding windows may overlap.

4.2.1.3 Example of Decomposition

Consider an SMF representing air quality across the state of Maine. A spatial partition by county yields a collection of sub-SMFs, each representing air quality measurements restricted to a single county. Alternatively, applying a tumbling window with a duration of one day yields a finite Measurement Field containing measurements from the most recent complete 24-hour interval.

4.2.2 Restriction

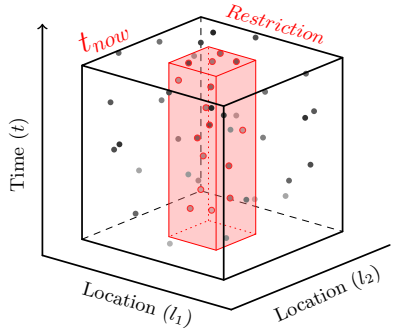


Figure 4 Spatial Restriction of SMF.

Restriction (Figure 4) limits the scope of an SMF by retaining only those measurements that satisfy a predicate. Restriction operates point-wise on measurements and preserves stream structure. Let σ be a predicate over individual measurements, parameterized by a dimension $D \in \{L, V, T\}$. Then:

$$\mathsf{Restrict}_{\sigma} : SMF \times \sigma \to SMF$$

Restriction can express point-wise temporal predicates (e.g., $v \geq 0$). However, interval-based temporal conditions defined relative to t_{now} (e.g., "last 5 minutes") require access to a bounded history and are therefore expressed via the *Window* operator introduced in the Decomposition section.

4.2.2.1 Example of Restriction

Consider an SMF representing temperature across the state of Maine. Let L_{York} denote the boundary of York County. Define the spatial predicate $\sigma_{York}(t, l, v) \equiv (l \in L_{York})$. Applying $\mathsf{Restrict}_{\sigma_{York}}$ yields an SMF containing only temperature measurements located within York County.

4.2.3 Combination

Combination operators (Figure 5) complement Decomposition and Restriction. While Decomposition separates an SMF along a chosen dimension and Restriction selects measurements satisfying predicates, Combination reconstitutes or relates compatible SMFs using binary set-theoretic operations. Together, these operators form a system for dividing, filtering, and recombining SMFs. Combination operates point-wise over time and produces a new SMF$'$ composed only of measurements already present in the operands. The general signature is:

$$\mathsf{Combine}_{\circ} : SMF \times SMF \to SMF,$$

where $\circ \in \{\cup, \cap, \setminus, \triangle\}$ denotes *union, intersection, difference*, or *symmetric difference*. Let SMF_A and SMF_B induce measurement sets M_A and M_B at time t_{now}. Each combination operator defines the resulting measurement set as $M' = M_A \circ M_B$.

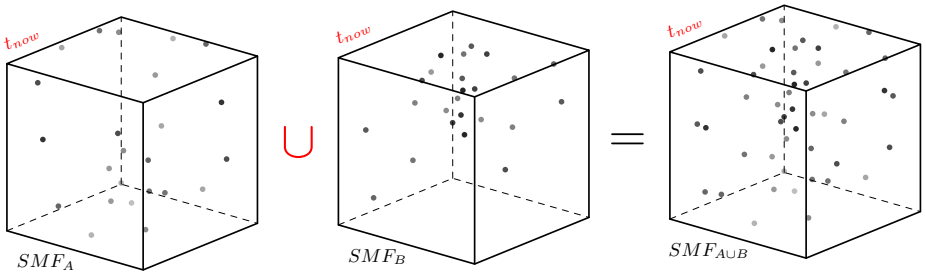


Figure 5 Union of two compatible SMFs.

These operators preserve stream structure and do not generate new measurements. Instead, they reorganize or select from existing measurements at each time index. Combination requires dimensional compatibility such that the two SMFs must share identical thematic definitions, compatible spatial reference systems, and synchronized temporal indices. If encodings differ (e.g., Celsius vs. Fahrenheit), a Transformation must be applied prior to combination. Combination is defined only for compatible theme and domain. If two SMFs contain competing values for the same property at the same location and time, the result is undefined until the conflict is reconciled through transformation, interpolation, summarization, or an application-specific rule.

4.2.3.1 Example of SMF Union

Let SMF_A represent air quality measurements in Orono, Maine, and SMF_B represent air quality measurements in Old Town, Maine. The union $SMF_A \cup SMF_B$ yields a new SMF representing air quality across both locations, preserving the temporal stream structure.

4.2.4 Numeric

Numeric operators derive new SMFs by applying primitive mathematical operations to values from an input SMF and a second operand. The second operand may be (i) another SMF, (ii) a static measurement field F (for example a historical snapshot), or (iii) a scalar constant. Numeric preserves stream progression and yields an SMF' whose measurements share the spatial and temporal identity of the left operand. Let $\circ \in \{+, -, \times, /\}$ denote a primitive numeric operator. We define the numeric operator class as:

$$\mathsf{Numeric}_\circ : SMF \times B \to SMF, \quad B \in \{SMF, F, \mathbb{R}\}.$$

At each time instant t_{now}, the left operand induces a measurement field $F_A(t_{now})$ and the right operand B induces a compatible value source. The base-case semantics assume one-to-one alignment by space-time identity. Thus, Numeric is defined for operands whose controls are aligned over $L \times T$. When supports differ, alignment must be established first through windowing, transformation, or interpolation. Formally, let $m_A = ((l,t), Z, v_A)$ be a measurement in the induced field of the left operand at t_{now}. If the right operand provides a corresponding value v_B for the same spacetime identity (l,t), or if the right operand is a scalar constant, $c \in \mathbb{R}$, then the resulting measurement is:

$$m' = \big((l,t), Z, v_A \circ v_B\big) \quad \text{or} \quad m' = \big((l,t), Z, v_A \circ c\big).$$

If no corresponding value exists in the right operand under identity alignment, the result at that spacetime identity is undefined and may be treated as missing. Missing data may be estimated using previously described interpolation. Note that in addition to local atomic operators, the Map Algebra uses focal, zonal, and global operators. In our framework, these may be approximated with decomposition and summarization such that the units of analysis within the operands are now regions rather than strictly atomic.

4.2.4.1 Example of Numeric between co-located SMFs

Let SMF_T be a temperature stream and SMF_H be a humidity stream produced by co-located sensors that report at identical spacetime coordinates. Define a numerical function ψ_{HI} that maps a temperature value and a humidity value to a heat-index value. The derived streamed field is:

$$SMF_{HI} = \mathsf{Numeric}_{\psi_{HI}}(SMF_{Temp}, SMF_{Hum}).$$

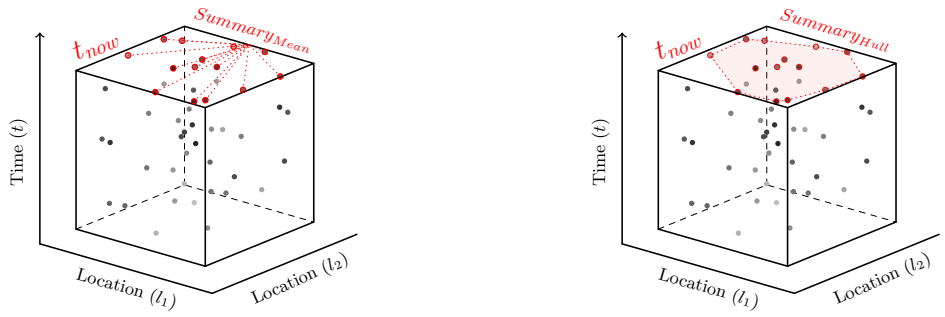
For each aligned pair of co-located and synchronized measurements

$$m_T = ((l,t), Z_{Temp}, v_{Temp}) \quad \text{and} \quad m_H = ((l,t), Z_{Hum}, v_{Hum}),$$

the resulting heat-index measurement is:

$$m_{HI} = \big((l,t), Z_{HI}, \psi_{HI}(v_{Temp}, v_{Hum})\big).$$

4.2.5 Summarization



(a) Metric Summary via Mean.

(b) Spatial Summary via Hull.

Figure 6 Summarization Operators.

Summarization collapses one or more dimensions of an SMF into a derived result that no longer retains the SMF structure. Let F_w be a finite measurement field and let ϕ specify the summary function and the dimensions to be collapsed. The operator has type

$$\mathsf{Summarize}_\phi : F_w \to R_\phi,$$

where R_ϕ is the result type determined by ϕ. For example, R_ϕ may be a scalar statistic, a time-indexed statistic, a region-indexed statistic, a geometry, or a structural descriptor. Summarization operates over the induced measurement field $F_{t_{now}}$, producing an aggregate, condensed, or geometric characterization of the measurements.

Value summarization collapses the metric dimension while retaining spatial or temporal context, yielding statistics such as mean(v), max(v), min(v), sum(v), or var(v) (Figure 6a). Location summarization collapses the spatial dimension, producing a set of locations satisfying a condition or a derived geometry such as a bounding box, convex hull, or alpha shape (Figure 6b).

Summarization may also collapse multiple dimensions jointly. Multidimensional summaries characterize relationships among values and their spatial or temporal configuration. Examples include spatial autocorrelation statistics such as Moran's I or Geary's C, which summarize similarity as a function of spatial proximity, and regression or spatiotemporal trend estimators that jointly summarize L, T, and V. In these cases, the result is a structural descriptor rather than a reduced subset of measurements.

4.2.5.1 Example of Summarization

Summarizing an SMF representing rainfall where $v > 50\,\text{mm}$ may yield a bounding region describing the spatial extent of heavy precipitation. This is distinct from restricting an SMF using an equivalent predicate which would preserve the structure and additional information within the SMF. Here we are collapsing the SMF to return locations.

4.3 Query Composition

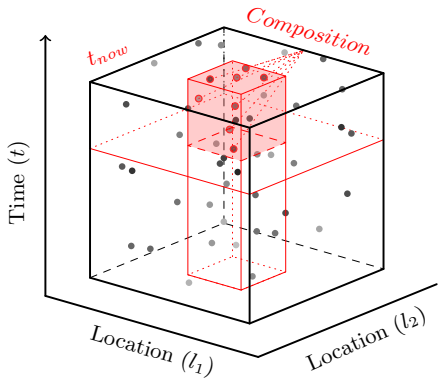


Figure 7 Workflow composition of Restriction, Windowing, and Summarization.

Query composition defines analytical workflows through the chaining of the operators. No new primitives are introduced; expressive power arises from combining the existing operator classes into executable pipelines under real-time constraints. Figure 7 illustrates a typical workflow:

$$\mathsf{Summarize}\big(\mathsf{Window}(\mathsf{Restrict}(SMF))\big).$$

In this example, the input SMF flows through three stages. First, *Restriction* reduces the spatial scope of the stream. Second, *Window* establishes a bounded temporal evaluation context. Finally, *Summarization* derives an aggregate descriptor from the finite measurement field produced by the window. The resulting output is the mean value of measurements within the selected region over the most recent complete window.

Composition therefore describes a directed evaluation workflow in which data are progressively transformed from streamed structure to bounded context to derived result. Operator ordering is governed by signature compatibility. In this workflow,

$$\mathsf{Restrict}_\sigma : SMF \to SMF, \quad \mathsf{Window}_{i,k} : SMF \to \{F_w\}, \quad \mathsf{Summarize}_\phi : F_w \to R_\phi.$$

Here, $\{F_w\}$ denotes the collection of finite measurement fields induced by the window specification, and each $F_w \in \{F_w\}$ may be passed to $\mathsf{Summarize}_\phi$.

Because $\mathsf{Summarize}_\phi$ collapses dimensional structure and produces a derived result rather than an SMF, it terminates the workflow. Execution order is thus determined by type compatibility and evaluation context.

5 Discussion

In this section, we validate our proposed operator set by comparing it against the requirements analysis in Section 3.3.1. Further, we also discuss some of its limitations.

5.1 Validation

We validate our analysis operator framework using two criteria: (i) *analytical expressiveness*, meaning systematic coverage of the recurring analytical behaviors for streamed spatiotemporal scalar fields identified in Section 3.3.1; and (ii) *minimality*, meaning that no operator class can be removed without losing essential analytical capability under forward-only evaluation and bounded state.

5.1.1 Analytical Expressiveness

Analytical expressiveness is established by explicitly matching each analytical behavior from Section 3.3.1 to either (a) a direct operator class defined in Section 4, or (b) a workflow expressed as a composition of the defined operator classes. The workflow patterns reuse the signatures and composition principles already introduced in Section 4 and Figure 7.

Behavior 1. The restriction of analytical scope is directly supported via the ***Restriction*** operator class. Spatial, temporal, and metric restriction are expressed as point-wise predicates over measurements (Section 4). Temporal restriction defined as an interval relative to t_{now} is realized as temporal decomposition through windowing (Behavior 2).

Behavior 2. The construction of finite evaluation units is directly supported through the ***Decomposition*** operator class. Spatial and thematic decomposition induce evaluation units such as regions, tessellations, or value classes, enabling tractable computation over unbounded streams. Temporal decomposition is realized through windowing, which bounds state relative to t_{now}: $\mathsf{Decompose}_T \equiv \mathsf{Window} : SMF \times W \to \{F\}$.

Behavior 3. The summarization of data can be interpreted globally or within units. Globally, this behavior is supported directly via the ***Summarization*** operator class. When summarization is to occur within sub-units of analysis, this behavior is supported through the composition of ***Summarization*** with ***Restriction*** or ***Decomposition***. For example, if we want to determine the mean temperature in Orono, Maine over the last hour every thirty minutes, we can compose a workflow utilizing behaviors 1, 2, and 3 as follows:

$$Temperature = \mathsf{Summarize}_{Mean}\Big(\mathsf{Decompose}_{T_{(60,30)}}\big(\mathsf{Restrict}_{L_{Orono}}(SMF)\big)\Big),$$

A single pipeline demonstrating the same task behaves as follows:

$$SMF \to \mathsf{Restrict}_{L_{Orono}} \to \mathsf{Decompose}_{T_{(60,30)}} \to \mathsf{Summarize}_{\mathrm{Mean}} \to Temperature$$

Behavior 4. Transformation of representation is directly supported via the ***Transformation*** operator class. This re-encodes spatial, temporal, or metric representations while preserving stream semantics. Transformation supports both standalone tasks (for representation compatibility) and pre-processing steps that enable downstream combination and comparison.

Behavior 5. Correlation between phenomena is expressed as a composition of ***Numeric*** and ***Summarization*** operators. Numeric operators construct derived and product fields, which are then aggregated within a bounded evaluation context. For two aligned streamed fields SMF_A and SMF_B, covariance may be expressed as:

$$\begin{array}{ccccccc}
SMF_A & \to & \mathsf{Summarize}_{\mathrm{Mean}} & \to & SMF_{\bar{A}} & & \\
& \searrow & & \swarrow & & & \\
& & \mathsf{Numeric}_{\mathrm{Minus}} & \to & SMF_{A'} & \searrow & \\
& & & & & & \mathsf{Numeric}_{\mathrm{Covariance}} \to R \\
& & \mathsf{Numeric}_{\mathrm{Minus}} & \to & SMF_{B'} & \nearrow & \\
& \nearrow & & \nwarrow & & & \\
SMF_B & \to & \mathsf{Summarize}_{\mathrm{Mean}} & \to & SMF_{\bar{B}} & &
\end{array}$$

Numeric provides the step that constructs interaction fields, while summarization performs bounded statistical aggregation. No additional primitive is required.

Behavior 6. Pattern and event recognition are expressed by a composition of ***Numeric*** and ***Summarization*** over bounded contexts. A threshold-based detector may be expressed as: $R = \mathsf{Summarize}_{\mathrm{DETECT}}\Big(\mathsf{Window}_W\big(\mathsf{Numeric}_\psi(SMF)\big)\Big)$, where ψ constructs binary or deviation fields prior to aggregation. Stream-feasibility is preserved through windowing.

Behavior 7. Numeric derivation of new fields is directly supported through ***Numeric***. This enables construction of derived fields such as difference fields, ratios, anomaly streams, interaction terms, and composite indicators.

Behavior 8. Neighborhood reasoning is expressed through a composition of spatial ***Decomposition*** followed by ***Numeric*** or ***Summarization***. Numeric may construct gradients or difference fields prior to summarization, enabling focal-style reasoning without introducing a dedicated neighborhood primitive.

5.1.2 Minimality of the Operator Set

Minimality is established at the level of operator classes under streamed constraints. An operator class is minimal if removing it eliminates expressiveness that cannot be reconstructed from the remaining classes without violating (i) bounded state, (ii) forward-only evaluation, or (iii) type preservation over the SMF. First we distinguish related operator classes.

Both *Restriction* and *Decomposition* operators limit analytical scope, but they do so with different type effects. *Restriction* preserves stream structure ($SMF \to SMF$) while applying predicates over time, location, or value. It reduces the count of $m \in SMF$ but still

yields a SMF. *Decomposition*, by contrast, constructs a collection of sub-SMFs or sets of finite Measurement Fields from unbounded streams ($SMF \to \{SMF\}$ or $\{F\}$). Although a decomposition may be implemented as a family of restrictions, it is not reducible to *Restriction* at the operator-class level because it returns $\{SMF\}$ or $\{F\}$ for subsequent analysis. Temporal decomposition, or windowing, is the mechanism that enforces bounded state relative to t_{now}, making rolling indicators well-defined.

Both *Numeric* and *Summarization* operators work over values, but they differ in type effect. *Numeric* operators generate new measurement values while preserving streamed structure ($SMF \to SMF$). Differences, ratios, anomaly streams, and interaction terms are value-level constructions that remain streams. *Summarization* collapses one or more dimensions to produce analytical descriptors ($SMF \to$ finite descriptor). Statistics, indicators, and structural measures require dimension reduction.

The remaining classes ensure structural and representational coherence in multi-input workflows. *Combination* enables structural recombination of compatible streamed fields. *Combination* operators reorganize compatible measurement sets. These are distinct operations at the type level. *Transformation* guarantees representational compatibility across spatial, temporal, or thematic encodings. Without it, semantically equivalent measurements encoded in different coordinate systems or units would be ill-typed under Combination or Numeric operators. *Interpolation* provides alignment under sparse or asynchronous sampling. Without it, multi-input workflows would implicitly assume perfect co-location and synchronization, which exceeds typical streamed measurement conditions. Taken together, these operator families form a minimal operator set: removing any family eliminates a capability that cannot be reconstructed from the others without violating streamed semantics or type discipline.

5.2 Limitations

The framework is deliberately constrained to analyses that remain well-defined under forward-only time and bounded state. Methods requiring unbounded historical access, backward temporal traversal, or end-of-stream knowledge fall outside the scope of stream-based analysis (e.g., full retrospective segmentation or long-horizon discovery over the entire past). Future work may introduce an additional architectural layer that integrates this framework with systems designed for long-term storage and retrospective time-series analysis; however, this paper focuses specifically on real-time stream processing over continuously evolving fields. The framework also distinguishes analysis from presentation and restructuring: operations such as pivoting, reshaping, or report-oriented schema transformations are important for visualization and communication, but they are not treated as core analytical tasks over continuous spatiotemporal phenomena. Object tracking, trajectory mining, geometric event streams, and human-sensed spatiotextual streams are adjacent geostreaming problems, but they fall outside this paper's scalar-field scope.

6 Related Work

In this paper, we have proposed a novel operator set that follows two criteria: it supports analysis of continuous spatiotemporal phenomena and it supports the computational rigor of stream processing.

Steenbergen et al. [40] define transformations among core geographic concept types such as objects, fields, networks, and events. For fields, the focus of this article, which they model conceptually as relations between locations and measured qualities, they introduce

representation-level operators such as interpolate for estimating a continuous field from point-wise samples, and invert or revert for switching between functional fields and coverage-based structures like contours. They also formalize field-level derivations, summaries and secondary fields. Spatiotemporal reasoning is not considered.

Decades ago, Map Algebra introduced the influential operator framework of local, focal, and zonal functions for deriving and summarizing fields through composition [41]. Cubic Map Algebra extended these ideas into spacetime [26]. Spatiotemporal algebras generalized field-based reasoning across additional dimensional configurations and transformations [15, 8]. These approaches make important contributions to the analysis of continuous phenomena and support fine-grained spatial and temporal reasoning. However, they generally assume materialized datasets in which time is retrospectively traversable, enabling global analysis over stored data. They therefore do not directly address forward-only evaluation, bounded memory, or window semantics characteristic of stream processing. The SMF occupies a restricted Sinton configuration in which space and time control a measured scalar theme, but adds stream-time semantics to this scalar-field setting. Within each window, the SMF supports the same kind of spatiotemporal field reasoning addressed by these prior algebras, while extending it to continuously arriving measurements under bounded-state streaming constraints.

By contrast, DSEs support low-latency processing of massive amounts of measurements and have introduced windowing, incremental aggregation and basic operators [5, 6]. Spatially enabled DSEs add spatial filtering over streams of spatial objects [44]. This work supports moving objects, spatial predicates, and vector geometries over streams [27, 21, 16, 44, 10]. Yet these systems operate over collections of low-level geometries and do not support higher-level abstractions such as fields or analytical operators for them.

Closest to our scope are approaches that treat sensor streams as observations of continuous phenomena, including window queries over continuous phenomena [42], phenomena-aware stream processing [3] and streamed field data types [24]. [43, 25, 3] focus on high-throughput query processing rather than a compact field-level operator vocabulary and proposed real-time algorithms for specific spatiotemporal field operations, such as interpolation or update propagation over high-volume streams [1, 23, 21, 25]. [24] introduced a higher-level abstraction for streamed fields that rely on the relational data model; however, it does not include analysis operators.

7 Conclusions and Future Work

This paper introduced a novel operator set for the analysis of stream-based spatiotemporal fields. While continuous geospatial phenomena are increasingly observed through dense, real-time measurement streams from geosensor networks and IoT platforms, data stream engines provide low-latency processing over unbounded streams, but they lack higher-level abstractions for spatially continuous phenomena. Our proposed analytics operator framework for streamed scalar fields builds on compact operator sets and dimensional reasoning over space, time, and theme. The resulting framework provides a systematic basis for describing and implementing extensible stream-based, real-time field analytics. Future work includes establishing the algebraic properties of these primitive operators for analyzing real-time continuous phenomena and implementing this design with a live geosensor network.

References

1 Md Mahbub Alam, Luis Torgo, and Albert Bifet. A survey on spatio-temporal data analytics systems. *ACM Comput. Surv.*, 54(10s), November 2022. doi:10.1145/3507904.

2 M. H. Ali, Mohamed F. Mokbel, and Walid G. Aref. Phenomenon-aware stream query processing. In *2007 International Conference on Mobile Data Management*, pages 8–15, 2007. doi:10.1109/MDM.2007.11.

3 Mohamed H. Ali. *Phenomenon-aware Data Stream Management Systems*. Phd dissertation, Purdue University, 2007. Accessed: 2026-05-27. URL: https://faculty.washington.edu/mhali/Publications/documents/ali_dissertation.pdf.

4 Arvind Arasu, Shivnath Babu, and Jennifer Widom. The cql continuous query language: semantic foundations and query execution. *The VLDB Journal*, 15(2):121–142, June 2006. doi:10.1007/s00778-004-0147-z.

5 Arvind Arasu and Jennifer Widom. A denotational semantics for continuous queries over streams and relations. *SIGMOD Rec.*, 33(3):6–11, September 2004. doi:10.1145/1031570.1031572.

6 Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. Association for Computing Machinery. doi:10.1145/543613.543615.

7 Shivnath Babu and Jennifer Widom. Continuous queries over data streams. *SIGMOD Rec.*, 30(3):109–120, September 2001. doi:10.1145/603867.603884.

8 Gilberto Camara, Max J. Egenhofer, Karine Ferreira, Pedro Andrade, Gilberto Queiroz, Alber Sanchez, Jim Jones, and Lubia Vinhas. Fields as a generic data type for big spatial data. In Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank, editors, *Geographic Information Science*, pages 159–172, Cham, 2014. Springer International Publishing. doi:10.1007/978-3-319-11593-1_11.

9 Donald D. Chamberlin and Raymond F. Boyce. Sequel: A structured english query language. In *Proceedings of the 1974 ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control*, SIGFIDET '74, pages 249–264, New York, NY, USA, 1974. Association for Computing Machinery. doi:10.1145/800296.811515.

10 Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Samuel R. Madden, Fred Reiss, and Mehul A. Shah. Telegraphcq: continuous dataflow processing. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, page 668, New York, NY, USA, 2003. Association for Computing Machinery. doi:10.1145/872757.872857.

11 E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387, June 1970. doi:10.1145/362384.362685.

12 George Colliat. Olap, relational, and multidimensional database systems. *SIGMOD Rec.*, 25(3):64–69, September 1996. doi:10.1145/234889.234901.

13 Open Geospatial Consortium. OGC API – Styles – Part 1: Core. OGC Draft Standard 19-087, Open Geospatial Consortium, 2019. Accessed 2026-05-27. URL: https://docs.ogc.org/DRAFTS/19-087.html.

14 Elias Dritsas and Maria Trigka. Remote sensing and geospatial analysis in the big data era: A survey. *Remote Sensing*, 17(3), 2025. doi:10.3390/rs17030550.

15 Karine Reis Ferreira, Gilberto Camara, and Antônio Miguel Vieira Monteiro. An algebra for spatiotemporal data: From observations to events. *Transactions in GIS*, 18(2):253–269, 2014. doi:10.1111/tgis.12030.

16 Z. Galić, M. Baranović, K. Križanović, and E. Mešković. Geospatial data streams: Formal framework and implementation. *Data and Knowledge Engineering*, 91:1–16, 2014. doi:10.1016/j.datak.2014.02.002.

17 Antony Galton. A formal theory of objects and fields. In *Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science*, COSIT 2001, pages 458–473, Berlin, Heidelberg, 2001. Springer-Verlag. doi:10.5555/646128.681085.

18 Antony Galton. Fields and objects in space, time, and space-time. *Spatial Cognition & Computation*, 4(1):39–68, 2004. doi:10.1207/s15427633scc0401_4.

19 Michael F. Goodchild, May Yuan, and Thomas J. Cova. Towards a general theory of geographic representation in gis. *Int. J. Geogr. Inf. Sci.*, 21(3):239–260, January 2007. doi:10.1080/13658810600965271.

20 Leila Hashemi Beni, Mir Abolfazl Mostafavi, Jacynthe Pouliot, and Marina Gavrilova. Toward 3d spatial dynamic field simulation within gis using kinetic voronoi diagram and delaunay tetrahedralization. *Int. J. Geogr. Inf. Sci.*, 25(1):25–50, February 2011. doi:10.1080/13658811003601430.

21 Seyed Jalal Kazemitabar, Ugur Demiryurek, Mohamed Ali, Afsin Akdogan, and Cyrus Shahabi. Geospatial stream query processing using microsoft sql server streaminsight. *Proc. VLDB Endow.*, 3(1–2):1537–1540, September 2010. doi:10.14778/1920841.1921032.

22 Karen K Kemp and Andrej Vckovsky. Towards an ontology of fields. In *International Conference on GeoComputation*, 1998. URL: https://www.geocomputation.org/1998/index.html.

23 Wenwen Li, Yan Liu, and Sizhe Wang. Real-time gis programming and geocomputation. *Geographic Information Science & Technology Body of Knowledge*, 2022, January 2022. doi:10.31223/X5CF22.

24 Qinghan Liang, Silvia Nittel, and Torsten Hahmann. From data streams to fields: Extending stream data models with field data types. In Jennifer A. Miller, David O'Sullivan, and Nancy Wiegand, editors, *Geographic Information Science*, pages 178–194, Cham, 2016. Springer International Publishing. doi:10.1007/978-3-319-45738-3_12.

25 Qinghan Liang, Silvia Nittel, John C. Whittier, and Sytze de Bruin. Real-time inverse distance weighting interpolation for streaming sensor data. *Transactions in GIS*, 22(5):1179–1204, 2018. doi:10.1111/tgis.12458.

26 Jeremy Mennis, Roland Viger, and C. Tomlin. Cubic map algebra functions for spatio-temporal analysis. *Cartography and Geographic Information Science*, 32:17–32, March 2013. doi:10.1559/1523040053270765.

27 Mohamed Mokbel, Xiaopeng Xiong, Moustafa Hammad, and Walid Aref. Continuous query processing of spatio-temporal data streams in place. *GeoInformatica*, 9:343–365, January 2005. doi:10.1007/s10707-005-4576-7.

28 Rohan Murty, Geoffrey Mainland, Ian Rose, Atanu Roy Chowdhury, Abhimanyu Gosain, Josh Bers, and Matt Welsh. Citysense: An urban-scale wireless sensor network and testbed. In *2008 IEEE Conference on Technologies for Homeland Security*, pages 583–588, June 2008. doi:10.1109/THS.2008.4534518.

29 Silvia Nittel. A survey of geosensor networks: Advances in dynamic environmental monitoring. *Sensors*, 9:5664–5678, July 2009. doi:10.3390/s90705664.

30 Silvia Nittel. Real-time sensor data streams. *SIGSPATIAL Special*, 7(2):22–28, September 2015. doi:10.1145/2826686.2826691.

31 Roger Cesarié Ntankouo Njila, Mir Abolfazl Mostafavi, and Jean Brodeur. A decentralized semantic reasoning approach for the detection and representation of continuous spatial dynamic phenomena in wireless sensor networks. *ISPRS International Journal of Geo-Information*, 10(3), 2021. doi:10.3390/ijgi10030182.

32 Ainhoa Osa Sanchez and Begoña Zapirain. Real-time air quality monitoring: A smart iot system using low-cost sensors and 3d printing. *IEEE Journal of Radio Frequency Identification*, PP:1–1, January 2025. doi:10.1109/JRFID.2025.3541816.

33 Yohanes Yohanie Fridelin Panduman, Nobuo Funabiki, Evianita Dewi Fajrianti, Shihao Fang, and Sritrusta Sukaridhoto. A survey of ai techniques in iot applications with use case investigations in the smart environmental monitoring and analytics in real-time iot platform. *Information*, 15(3), 2024. doi:10.3390/info15030153.

34 Jinho Park, Taeyoung Yoo, Seongjae Lee, and Taehyoun Kim. Urban noise analysis and emergency detection system using lightweight end-to-end convolutional neural network. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, 18, August 2023. doi:10.15837/ijccc.2023.5.5814.

35 JJ Paul and Silvia Nittel. The streamed measurement field model: Supporting real-time geospatial analysis of continuous phenomena using streams. *Under Review*, 2026.

36 Loïc Petit, Cyril Labbé, and Claudia Lucia Roncancio. An algebric window model for data stream management. In *Proceedings of the Ninth ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '10, pages 17–24, New York, NY, USA, 2010. Association for Computing Machinery. doi:10.1145/1850822.1850826.

37 Donna J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994. doi:10.1111/j.1467-8306.1994.tb01869.x.

38 T. Poston. *Fuzzy geometry*. Phd dissertation, University of Warwick, June 1971. URL: http://webcat.warwick.ac.uk/record=b1733602~S1.

39 David S. Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard Papers on Geographic Information Systems*, 6:1–17, 1978.

40 Niels Steenbergen, Eric Top, Enkhbold Nyamsuren, and Simon Scheider. Procedural metadata for geographic information using an algebra of core concept transformations. *Journal of Spatial Information Science*, pages 51–92, December 2023. doi:10.5311/JOSIS.2023.27.268.

41 C. Dana Tomlin. Map algebra: one perspective. *Landscape and Urban Planning*, 30(1):3–12, 1994. Special Issue Landscape Planning: Expanding the Tool Kit. doi:10.1016/0169-2046(94)90063-9.

42 J. C. Whittier, Silvia Nittel, Mark A. Plummer, and Qinghan Liang. Towards window stream queries over continuous phenomena. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '13, pages 2–11, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2534303.2534305.

43 J.C. Whittier. *Towards an Efficient, Scalable Stream Query Operator Framework for Representing and Analyzing Continuous Spatio-Temporal Fields*. Dissertation, University of Maine, 2018. URL: https://digitalcommons.library.umaine.edu/etd/2927.

44 Jia Yu, Jinxuan Wu, and Mohamed Sarwat. Geospark: a cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2820783.2820860.

45 Yuanzhen Zhang, Guofang Wang, Lingzhi Li, and Mingjing Huang. A monitoring method for agricultural soil moisture using wireless sensors and the biswas model. *Agriculture (Switzerland)*, 15, February 2025. doi:10.3390/agriculture15030344.

46 Ariane Ziehn, Jan Szlang, Steffen Zeuch, and Volker Markl. Unraveling the impact of window semantics: Optimizing join order for efficient stream processing. In *Proceedings of the VLDB Endowment*, volume 18, pages 2468–2481. VLDB Endowment, 2025. doi:10.14778/3742728.3742741.

Through and Through: Corpus-Based Semantic Analysis of a Spatial Preposition

Brandon Bennett
School of Computing, University of Leeds, UK

Maria M. Hedblom
Department of Computer and Systems Sciences, Stockholm University, Sweden
Department of Computing, Jönköping University, Sweden

Abstract

We present a corpus-based analysis of sentences containing the highly polysemous word "through" and establish a framework for categorising sentence constituents in relation to this preposition. Using the annotated corpus, together with ideas from the fields of Lexical Semantics and Knowledge Representation, we specify a typology of *figure*, *path-verb*, and *ground* terms, characterised by the *affordances* they provide for participation in "through" relationships. We find that the mode of affordance provided by the *ground* is often ambiguous: many types of object can provide multiple modes of grounding for a "through" relationship, associated with different geometrical conditions between figure and ground. However, this ambiguity can be explained and often resolved or reduced by considering compatible combinations of figure and ground affordances in relation to the path.

2012 ACM Subject Classification Computing methodologies → Knowledge representation and reasoning; Computing methodologies → Natural language processing; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Preposition, Polysemy, Semantics, Spatial Relation, Ontology

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.15

Category Short Paper

1 Introduction

Prepositions are an essential component of English and many other natural languages. They form a relatively small ("closed") class of words that combine with other words to describe fundamental spatial and temporal relationships between objects in the word. A preposition that displays a very wide range of senses is "through." While usually regarded as primarily a spatial preposition, it has many uses in describing temporal, causal, informational and other abstract relationships. It can occur in several grammatical roles and interacts in complex ways with associated verbs as well as both physical and abstract objects. Even when confining our attention to spatial descriptions, we can contrast cases such as: *"The boy ran through the tunnel"*, *"A rope ran through the tunnel"* and *"Water ran through the tunnel"*; as well as *"She ran a comb through her hair"*, *"She threw the comb to him through the window"* and *"Moonlight shone through the window."*

One may feel that there is some essential "throughness" that unites the situations. According to the *Image Schema* approach, the senses of prepositions, correspond to fundamental but highly adaptable patterns in our experience. [7, 5]. Each use of "through" conveys a particular range of possible spatio-temporal relationships holding between the subject (the *figure*) and the prepositional object (the *ground*) but the geometrical relationships involved can be very different. We aim to illuminate this diversity by analysing a corpus of sentences containing "through" and developing a framework for describing the roles of sentence constituents in such constructions. Drawing on ideas from lexical semantics and knowledge representation, we introduce a typology of *figure*, *ground*, and *path-verb* expressions, characterised by the affordances they contribute to a through relationship.

© Brandon Bennett and Maria M. Hedblom;
licensed under Creative Commons License CC-BY 4.0

2 Foundations and Previous Work

The massive variety of meanings conveyed by the word "through" offers an interesting study of the phenomenon of *polysemy*, where a word can be used in many different but related senses [4]. Because "through" can be used in so many different ways, it is difficult to determine whether two different examples can be considered as two cases of a general form of situation, or whether they exemplify different senses (See Section 4). Hence, we felt the need for an organising framework. The influential work of Talmy [9] construes the semantics of motion verbs and event descriptions as decomposable into the following interacting components:

Figure: the entity whose location or configuration changes,

Ground: the reference entity with respect to which the change is characterised,

Motion: the occurrence of displacement or maintenance of location over time,

Path: the geometric or topological structure of that displacement (direction, boundary-crossing, traversal),

Manner: the way in which the Motion is carried out (running, sliding, limping, crawling),

Cause: whether the Motion is externally induced (push, throw, carry).

In English Motion and often Manner are encoded in the verb, while properties of the Path are expressed by a preposition or particle (e.g. *run into*, *fall off*). For present purposes, this division is central, since the preposition "through" contributes a specification of Path that interacts systematically with the Figure, the Ground, and the verb.

A complementary perspective is provided by Pustejovsky's *Generative Lexicon* theory [8]. Rather than treating polysemy as a fixed list of discrete senses, the theory proposes that lexical items encode rich internal structure from which context-sensitive interpretations are generated. Composition is guided by a system of types and by mechanisms of coercion that allow operator expressions (verbs, adjectives, prepositions) to adjust the semantic type of their arguments when required. Hence, the meaning of a construction of the form "α ϕ's 'through' β" is not obtained by simply selecting a stored sense of "through", but emerges from interaction between (i) the semantic type and internal structure of the Figure α, (ii) the event structure and path-contributing properties of the verb ϕ, (iii) the structural and topological properties of the Ground β, and (iv) a family of constraints associated with "through". The resulting interpretation is the outcome of a compatibility resolution among these elements.

A central component of the Generative Lexicon is the *qualia structure*, which characterises an object along four dimensions: *formal* (what kind of thing it is), *constitutive* (its material or parts), *telic* (its function or purpose), and *agentive* (how it comes into being). For spatial expressions, such structure is particularly relevant because it encodes affordances: a tunnel affords passage, a wall resists it, a net permits partial penetration while constraining motion. Many uses of "through" appear to depend not merely on geometry but on such functionally grounded properties of the Ground and Figure. The qualia framework therefore, provides a principled way of modelling why certain objects are naturally construed as passable, penetrable, or traversable.

3 Our Methodology

Our methodology combines intuitive judgements about possible meanings of the word "through" with formal analysis from Lexical Semantics and logical theories of spatial and temporal relations. For our study we used the fiction texts of the Brown Corpus [6],[1] a widely

[1] The sentences of the corpus, the Python code used in the analysis, the semantic annotations, and some simple functions for displaying annotated sentences, category structures and frequency counts are available via *GitHub* (made available on acceptance to maintain anonymity).

used dataset suitable for reproducible analysis. The fiction section contains 275 occurrences of the word "through". Although not comprehensive, the relatively small but spatially rich sample allows manual annotation of every instance.

For each occurrence we recorded (i) the associated verb, (ii) the figure, the entity whose participation in the relation is conceptually central, and (iii) the ground, the entity or region introduced by "through". This stage required careful hand annotation because syntactic position was often misleading: the figure is not always the grammatical subject and may sometimes be implicit. In parallel we classified grounds into salient semantic types (see Section 4.3) and began a corresponding classification of verbs into event types (see Table 2).

Lightweight Python scripts were used for annotation and analysis but not for the semantic analysis itself. A language model (**ChatGPT-5.2**) was also used to assist with coding and occasionally to suggest likely sentence components or provisional classifications. Its output was treated only as indicative and was always checked against the original sentences.

Rather than beginning with a fixed ontology, we moved iteratively between the corpus and the developing classification. New examples often prompted additions or refinements to the category system. At the same time we sought generality and definability, identifying characteristics that applied across many cases and appeared amenable to precise definition. In this way corpus evidence, linguistic intuition, and type-based reasoning were used together, producing an analysis that is both empirically grounded and systematically organised.

4 Analysis

One may initially consider "through" to be a binary spatial relation. However, consideration of examples of its use reveals that its meaning often depends on an associated verb. Hence, the expression of "through" relationship can be decomposed into the following elements:

Figure, *Path-Verb*, ***through***, *Ground*.

4.1 Figures and their Types

Many types of entities can be the Figure in a "through" relationship description (see Table 1). Different types of physical objects provide different affordances, which influence the ways they can participate in "through" relationships. Hence, we have subcategorised PHYSICAL-OBJECT with respect to salient affordances as follows: ELONGATED-OBJECT ("sticks", "rifle muzzles", "bolo", "stake"), ANATOMICAL-OBJECT ("skull", "hair", "body"), PROJECTILE ("bullet", and two pronouns with "bullet" implied), AGGREGATE ("snowflakes"), PHYSICAL-PATH ("alleyway"), BUILDING ("cottages"), LIQUID-OBJECT ("river"), FILAMENTED-INSTRUMENT ("comb"), PHYSICAL-OBJECT ("basket"), PHYSICAL-BARRIER ("revetments"), *Mixed-type Figure* ("guns and fists").

The Figure contributes its trajectory (whose properties may also be described by the verb) and properties such as self-motivation, elongation, fluidity, hardness, or sharpness.

4.2 Verbs Associated with "Through"

We examined the verbs occurring with through in every sentence of the corpus and grouped them into categories (Table 2). Our categorisation is intended to cover all categories in which the verb can function. For instance, "run through" appears in the corpus only in the sense we call ITERATIVE-SCAN (as in "run through the info"). However, run has several senses, and is therefore also included in the SELF-MOTION, STATIC-ROUTE, and CAUSED-MOTION categories.

Table 1 Figure types in the "through" corpus.

Type	Count	Sub-Types
PERSON	174	Includes plurals, 11 cases implied/unreferenced.
PHYSICAL-OBJECT	20	'sticks', 'hair', 'cottages'
ANIMAL	5	3 'horse', 'nymph' (insect), 'birds'
VEHICLE	4	'jeep', 'van', 'truck', 'wagon'
LIGHT	11	5 'light', 2 daylight, 'moonlight', 'colours', ...
INFORMATION	10	'lyrics', 'news', 'messages', 'smile', 'images', ...
EMOTION/SENSATION	8	'twinges', 'loneliness', 'pity', 'chill', 'anger', ...
MENTAL-ENTITY/CAPACITY	4	3 'mind', 'thought'
GASEOUS-SUBSTANCE	8	2 'air', '2 'smoke', 'scent', 'breath', 'gases', 'wind'
LIQUID-SUBSTANCE	5	5 'water', 'sweat'
EVENT/PROCESS	5	'life', 'flight', 'childhood', 'deal', ..
SOUND	5	4 'sound', 'voice'
Other types with only 1 example	7	SOLID-SUBSTANCE, PHYSICAL-MOVEMENT, COLOUR, PLACE, GEOM-FIGURE, AUDIO-QUALITY, IMPLIED-GOAL

Table 2 Verb types with respect to *through* relationships. Verbs may appear in multiple categories due to polysemy.

Verb Type	Count	Items
SELF-MOTION	114	bolt, climb, come, crawl, cross, disappear, dive, drive, ...
STATIC-ROUTE	43	go, lead, lie, pass, run, thread
TRANSMISSION	36	carry, come, enter, flash, glow, pass, reach, show, ...
COMPLETION	36	come, get, go
PROPAGATION	28	course, disappear, filter, flood, glow, ripple, run, seep, ...
PERCEPTION	28	gaze, glance, hear, look, peek, peer, read, see, watch
CAUSED-MOTION	22	bring, carry, coax, drag, draw, force, hand, hurl, puff, ...
PENETRATION	17	crash, cut, drive, jab, pierce, rake, scuff, shoot, tear, wear
ITERATIVE-SCAN	15	leaf, pick, plod, run, search, seek, sift
ENDURANCE	8	carry, live, nurse, stay, suffer, sustain
CAUSAL-MEDIATION	5	acquaint, do, know, smooth, start

4.3 Ground Descriptors and their Types

We found great variety in affordances supplied by the Ground in relation to "through" and found them much harder to describe and categorise than those associated with the Figure. It was often difficult to identify genuinely different modes of affordance among the many possible geometries that an instance "through" relation may take. By looking for conceptual generalities among the particular cases we arived at the following typology:

- CONTAINER – an enclosed region that can be entered and exited.
  - MULTI-APERTURE – multiple distinct openings enabling traversal across it.
  - SINGLE-APERTURE – essentially one entry/exit opening.
  - STRUCTURED – internally partitioned into compartments.
- FRAME – a bounded opening defined by a surrounding object rather than a volume.
- APERTURE – an opening in a barrier separating two sides.
- APERTURE-PROXY – an object type used to refer to an associated aperture (e.g. a door or curtain).

- BARRIER – a surface or sheet obstructing movement between regions.
  - GAPPY-BARRIER – barrier containing holes or spaces permitting passage.
- AGGREGATE – a collection of discrete objects forming an extended obstruction or medium.
- ENVIRONMENT – an extended surrounding spatial region.
  - ENV-2.5D – surface-dominated terrain with horizontal extent.
  - ENV-3D – volumetric enveloping medium (air, fog, water, darkness).
  - CONDUIT – elongated environmental channel guiding motion (street, canyon).
- INFORMATION-CONDUIT – supports communication or perceptual access.
- BODILY-MEDIUM – parts of an organism functioning as a penetrable or perceptual medium.
- TIME-PERIOD – an extended temporal interval conceptualised as traversable.
- ENACTMENT – a sequence of actions forming a coherent episode.
- CAUSAL-MEDIUM – an abstract condition or circumstance producing an outcome.
- INFORMATION-SOURCE – a structured body of information examined sequentially.
- AUDIO-LAYER – a surrounding field of sound.
- FIGURE-2D – a purely geometric region.
- Additional Qualifier Types: UPRIGHT, GROUNDED (anchored to the ground), FILAMENTED (grass, fingers), LIGHT-TRANSMITTING, LOS-TRANSMITTING (apertures affording a line of sight).

4.3.1 Ground Vocabulary Type Lexicon

We use the terms above to define an affordance lexicon, which associates nouns occurring as the Ground in a "through" phrase with a boolean combination of categories and qualifiers. Of particular interest are those whose types are *disjunctive* (with $\vee$ in the type specification). These are cases where the ground object presents more than one mode of affordance with respect to potential "through" relationships. We found that anatomical objects provide especially interesting diversity in their affordances.

Informal grouping identified families of ground types: Containers, Apertures, Barriers and aggregates, Vegetated environments, Environment situated conduits, 3D environments, Anatomical objects, Time and Informational media. Each of these families provides a distinct range of affordance types in association with "through"; and, moreover, even these refined types could often provide several alternative affordances depending on their context. Complex affordance types are represented as follows:

Barriers and aggregates

grille, veil	GAPPY-BARRIER
curtain	BARRIER $\vee$ APERTURE-PROXY
fence	(GROUNDED $\wedge$ UPRIGHT $\wedge$ BARRIER) $\vee$ (GROUNDED $\wedge$ UPRIGHT $\wedge$ GAPPY-BARRIER)
trees	FILAMENTED $\vee$ (GROUNDED $\wedge$ UPRIGHT $\wedge$ AGGREGATE)
crowd, traffic, grass	GROUNDED $\wedge$ UPRIGHT $\wedge$ AGGREGATE

Anatomical Objects (often construed as Bodily Media)

body	BODILY-MEDIUM
hand	BODILY-MEDIUM $\wedge$ FILAMENTED
finger	BODILY-MEDIUM $\wedge$ FILAMENTED $\wedge$ INFORMATION-CONDUIT
nose	BODILY-MEDIUM $\wedge$ MULTI-APERTURE $\wedge$ CONTAINER
eye	BODILY-MEDIUM $\wedge$ LOS-TRANSMITTING $\wedge$ INFORMATION-CONDUIT

Typically, the *head* noun of a phrase yields the affordances of its referent. But this is not always the case. For example, "opaque glass" lacks the light-transmission property normally afforded by "glass". Hence, we use an additional phrase-level lexicon, so that affordances of objects described by multi-word phrases are more accurately captured.

5 Sense Annotations and Affordance Combinations

We now present some of the outputs from the main display script that we use to view and evaluate our annotations. To give a general flavour, we start by just giving the first two, and give one more example that we consider to be of interest.

```
0. "Rachel wore a smart hat and, because she had been warned recently about
    smoking, puffed at her cigarettes *through* a long ivory holder stained
    with lipstick."
       FIGURE: 'she' (PERSON)
    PATH VERB: 'puffed' (puff) (CAUSED-MOTION)
       GROUND: 'holder' ((SINGLE-APERTURE | MULTI-APERTURE) CONTAINER)

1. "At 2130 hours they had passed *through* the barbed wire at the point
    of departure."
       FIGURE: 'they' (PERSON)
    PATH VERB: 'passed' (pass) (TRANSMISSION | SELF-MOTION | STATIC-ROUTE)
       GROUND: 'barbed wire' (GROUNDED UPRIGHT FILAMENT BARRIER)

6. "He thought how this dainty, fragile older woman threading her way *through*
    the streets of Westminster on a day in June, enjoying the flowers in the
    shops, the greetings from old friends, but never really drawing a deep,
    passionate breath, was so like himself."
       FIGURE: 'this dainty, fragile older woman' (PERSON)
    PATH VERB: 'threading' (thread) (SELF-MOTION)
       GROUND: 'day' (TIME-PERIOD)
```

The final example is interesting because of the error in the *ground* annotation, which we subsequently corrected. Here, the error is plain to see because the affordances of the *figure* and *verb* do not fit with those of the *ground*.

5.1 Distinct Spatio-Temporal Relationships

We now turn to the question of what are the spatial conditions described by a "through" relationship. We are currently working with the following typology of "through" relations:

- APERTURE-CROSSING: Passage through a local opening in an otherwise blocking surface. (e.g. through the door, window, gap, keyhole)
- APERTURE-TO-APERTURE-TRAVERSAL: Passage through a local opening in an otherwise blocking surface. (e.g. through the door, window, gap, keyhole)
- CONDUIT-TRAVERSAL: Guided movement within an extended enclosed channel. (e.g. through the tunnel, pipe, cave passage)
- DESTRUCTIVE-BARRIER-PENETRATION: Penetration of resistant barrier (e.g. bullet through the wall, nail through the board, knife through fabric, brick through windowpane) Salient sub-types are: PIERCING and SMASHING/BREAKING
- CONTINUOUS-MEDIUM-TRAVERSAL: Motion within a yielding surrounding substance. (e.g. through water, air)
- AGGREGATE-OBJECT-TRAVERSAL: Passage through a densely occupied region. (e.g. through the crowd, traffic, forest)
- AGGREGATE-BARRIER-CROSSING: Crossing a barrier constituted by an aggregate.
- INTERIOR-REGION-TRAVERSAL: Movement across a bounded bounded volume. (e.g. through the room, building, city)

- THREADED-TRAVERSAL: Guided passage along a shaped interior path. (e.g. thread through the needle, rope through the pulley)

Our ongoing work is targeting two goals: (1) defining the geometrical conditions of spatio-temporal relationships corresponding to the different senses of "through"; (2) showing how our typology of the Figure, Verb and Ground types can be used to resolve or significantly reduce the ambiguity of "through" in any given context; and hence determine the spatio-temporal conditions it implies. Though we do not have space to present them here, we have already specified representations of several of the above relations in a spatio-temporal logic, making use of some ideas presented in [2] and [3], and of similar style to that presented in [1]. From our annotation outputs (see Section 5) we have worked out mappings to spatial conditions for many cases in our corpus.

6 Conclusion

Due to its huge range of senses, the meaning of "through" is resistant to straightforward classification and formalisation. Nevertheless, we believe that examining deeper semantic structure in terms of types and affordances of *figure*, *path-verb* and *ground* provides a promising route to a systematic account, that could support resolution of sense ambiguities of this highly polysemic word over a wide range of contexts. Although focused on a single word, we believe that our meticulous categorisation methodology can be applied to other words, and and may shed light on the phenomena of polysemy. Our preliminary findings are consistent with the view that meanings of polysemic words are resolved by accommodation of affordances between interacting semantic components within a sentence. Specifying mechanisms for this resolution is a remaining aim for the further development of our analysis.

References

1 Brandon Bennett, Vinay Chaudhri, and Nikhil Dinesh. A vocabulary of topological and containment relations for a practical biological ontology. In John Stell, Thora Tenbrink, and Zena Wood, editors, *Spatial Information Theory, 11th International Conference, COSIT 2013*, LNCS, Scarborough, UK, 2013. Springer. doi:10.1007/978-3-319-01790-7_23.

2 Anthony G Cohn, Brandon Bennett, John M Gooday, and Nick Gotts. Rcc: a calculus for region based qualitative spatial reasoning. *GeoInformatica*, 1(3):275–316, 1997.

3 Nicholas Mark Gotts. How far can we 'C'? Defining a 'doughnut'using connection alone. In *Principles of Knowledge Representation and Reasoning*, pages 246–257. Elsevier, 1994.

4 Stefan Th. Gries. Polysemy. *Cognitive linguistics: Key topics*, pages 473–490, 2019.

5 Maria M. Hedblom, Dagmar Gromann, and Oliver Kutz. In, Out and Through: Formalising some dynamic aspects of the image schema Containment. In *Proc. 33rd Annual ACM Symposium on Applied Computing*, pages 918–925, Pau, France, 2018. doi:10.1145/3167132.3167233.

6 Henry Kučera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, 1967.

7 George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

8 James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

9 Leonard Talmy. *Toward a Cognitive Semantics*. MIT Press, Cambridge, MA, 2000.

From Trajectories to Relations: Interpreting Overtake Manoeuvres with Large Language Models and Qualitative Spatial Representations

Jana Verdoodt
Department of Geography, Ghent University, Belgium
GeoAI Research Center, Ghent University, Belgium

Changbo Zhang
Department of Geography, Ghent University, Belgium

Lars De Sloover
Department of Geography, Ghent University, Belgium
GeoAI Research Center, Ghent University, Belgium

Haosheng Huang
Department of Geography, Ghent University, Belgium
GeoAI Research Center, Ghent University, Belgium

Nico Van de Weghe
Department of Geography, Ghent University, Belgium
GeoAI Research Center, Ghent University, Belgium

Abstract

This paper explores whether qualitative spatial encodings make overtake manoeuvres more interpretable to a large language model (LLM). Using simulated traffic trajectories from a microscopic road-network model, we encode the same overtake event at three levels of complexity through Point-Descriptor-Precedence (PDP) inequality matrices. An LLM is then used in a staged, zero-shot API workflow in which longitudinal and lateral descriptor encodings derived from these matrices are interpreted separately and then merged into a final manoeuvre interpretation. Outputs are assessed through expert-based qualitative evaluation focusing on relational correctness, spatial precision, completeness, and interpretive adequacy. For the selected case, the results suggest that the model can produce meaningful natural-language interpretations at all three levels, but that these change as complexity increases. The 2-point representation yields the clearest focal manoeuvre, the 5-point representation offers the best balance between contextual richness and interpretability, and the 10-point representation produces the most nuanced structural reading, although in a less concise form. The paper should be read as an exploratory study of representation effects rather than as a benchmark of LLM performance.

2012 ACM Subject Classification Computing methodologies → Natural language generation; Computing methodologies → Spatial and physical reasoning

Keywords and phrases micro-scale traffic analysis, overtaking behaviour, symbolic movement encoding, spatial reasoning, expert evaluation

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.16

Category Short Paper

Supplementary Material *Dataset (Derived qualitative encodings, API materials, visualisations, and expert evaluations)*: `https://doi.org/10.5281/zenodo.20729343`

Acknowledgements The authors thank MINT nv for providing the simulated traffic trajectory data used in this study.

© Jana Verdoodt, Changbo Zhang, Lars De Sloover, Haosheng Huang, and Nico Van de Weghe;
licensed under Creative Commons License CC-BY 4.0

1 Introduction

How can a movement event be represented so that its spatial structure can be interpreted, rather than merely measured? This question lies at the heart of qualitative spatial reasoning (QSR), where the aim is not only to record coordinates, but to describe relations such as before, behind, left of, or alongside in a form that supports spatial reasoning [6, 3]. This relational perspective is especially relevant for traffic manoeuvres such as overtaking. An overtake is not simply the path of one vehicle, but a structured change in longitudinal and lateral relations between a focal vehicle and surrounding traffic, making it a useful event for studying qualitative movement representation [15].

Detailed trajectory data are increasingly available through video-based observation, LiDAR, drones, and automated tracking [2, 18]. Most traffic analyses nevertheless treat manoeuvres primarily as numerical patterns to cluster, predict, or score. Prior work in qualitative movement representation and micro-scale traffic analysis has shown that vehicle interactions can also be treated as structured relational events, where subtle changes in position, ordering, and grouping may be difficult to detect from raw trajectories or aggregate safety indicators alone [14, 16, 10, 12]. In parallel, recent work has evaluated the spatial reasoning abilities of large language models (LLMs) and large reasoning models, including reasoning over RCC-8 relations, cardinal directions, StepGame, and broader spatial concepts [4, 5, 8, 9, 7, 17]. However, these studies mainly rely on synthetic or text-based inputs, rather than qualitative encodings derived from movement data.

In this short paper, we ask whether an LLM can coherently interpret a qualitative spatial encoding of the same overtake manoeuvre, and how that interpretation changes as the relational context becomes more complex. Using simulated traffic trajectories from a microscopic road-network model, we represent one focal overtake event at three levels of complexity through Point-Descriptor-Precedence (PDP) inequality matrices [11]. Rather than benchmarking LLM performance at scale, we examine how representation complexity affects the interpretation of a single, controlled manoeuvre case.

2 Data and Methods

Figure 1 presents the workflow used in this study: selected overtake cases are encoded through the Point-Descriptor-Precedence (PDP) framework [11], represented as qualitative inequality matrices, interpreted by an LLM, and then assessed through expert-based qualitative evaluation.

Figure 1 Overview of the workflow used in this study.

2.1 Traffic data and illustrative case selection

This study uses trajectory-based traffic data provided by an independent traffic and mobility consultancy. The data originate from a PTV VISSIM microscopic simulation of a three-lane highway environment, informed by representative traffic counts and regional origin–destination

matrices for the base year 2019–2020. From approximately 120 overtaking configurations available for each representation level, we focus on configuration 52 because it provides a visually clear and qualitatively well-bounded example of the same overtaking event with 2, 5, and 10 moving points. The three representation levels encode the same focal overtaking manoeuvre rather than three different traffic events. The 2-point representation contains only the two vehicles directly involved in the overtake and therefore represents the minimal focal interaction. The 5-point representation adds three surrounding vehicles to capture the immediate local traffic context, while the 10-point representation includes eight surrounding vehicles and serves as a richer complexity test. At each time step, the number of unique object pairs increases from 1 in the 2-point case to 10 and 45 in the 5- and 10-point cases. Across the full time window, the matrices also include relations between point instances at different time steps.

2.2 PDP encoding of overtake manoeuvres

To make the trajectories suitable for spatiotemporal reasoning, we use a qualitative relational representation derived from the PDP framework (Figure 2). In this representation, moving objects are expressed as points relative to selected descriptors, and their relative positions are encoded in a qualitative rather than metric form [11]. We apply the "rough" parameter only to the lateral descriptor, using a value of 1, so that small lateral coordinate differences within the same lane are treated as qualitatively equal rather than as distinct positions [10]. This prevents vehicles travelling in the same lane from being encoded as if they occupied different lateral positions solely because their coordinates are not exactly identical.

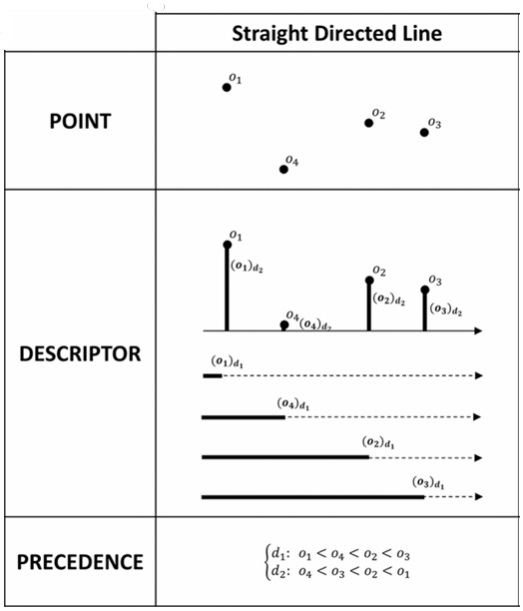


Figure 2 Conceptual illustration of the Point-Descriptor-Precedence (PDP) framework [10].

2.3 Qualitative inequality matrices

The PDP encoding results in inequality matrices that capture pairwise relations between points and provide the basis for the qualitative encodings submitted to the LLM. Separate matrices are constructed for the longitudinal x-descriptor and lateral y-descriptor. Pairwise

relations are encoded numerically, with 0, 1, and 2 corresponding to the qualitative PDP relations $<$, $=$, and $>$, respectively. Along the longitudinal descriptor, these indicate behind, equal, and ahead; along the lateral descriptor, they indicate one side, equal lane-level position, and the other side, relative to the descriptor orientation. Each matrix represents pairwise qualitative relations across a fixed time window of 50 steps covering the selected overtaking event. Increasing the number of moving points adds contextual detail but also increases representational complexity, allowing us to examine how LLM reasoning changes as the spatial encoding becomes more detailed. For expert inspection, the matrices were also rendered as colour-coded heatmaps showing the qualitative PDP relations.

2.4 LLM interpretation procedure

We evaluated LLM-based interpretation through an instruction-based zero-shot API workflow. For each representation, the model was queried with structured qualitative encodings derived from the inequality matrices rather than raw trajectory coordinates or visualisations. The longitudinal and lateral descriptor encodings were first submitted separately, after which their partial interpretations were merged into a final manoeuvre interpretation. Each descriptor-specific prompt asked the model to interpret changes in pairwise relations over time using only the provided qualitative encoding. A final prompt then merged both outputs into a structured description of the manoeuvre, its phases, and the roles of the involved points. All calls used GPT-5.4 through the API with fixed generation settings across runs (`temperature` = 0 and `top_p` = 1). Inputs were structured JSON payloads containing configuration metadata, encoding definitions, and compressed same-time point-pair summaries derived from the qualitative inequality matrices. The complete wording of the prompts, together with the JSON schema, API settings, and representative request–response examples, is provided in the supplementary materials to support reproducibility.

2.5 Expert-based qualitative evaluation

Model outputs were assessed through a two-expert qualitative evaluation. Independently for each representation level, both experts first established a reference interpretation using the static visualisation, animation, and inequality matrices or heatmaps, without consulting the LLM output. The merged LLM interpretation was subsequently compared with these expert references. Evaluation focused on relational correctness, spatial precision, completeness, and interpretive adequacy, using the categories “Strong”, “Adequate”, “Partial”, and “Weak”. The aim was not to benchmark LLM performance quantitatively, but to assess whether the model reconstructed the main relational structure in a spatially coherent and encoding-grounded way. The expert template follows broader work on human evaluation of LLM outputs [1, 13].

3 Results

3.1 A baseline overtake case and its contextual extensions

Figure 3 shows our configuration at three levels of representational complexity. Across the three representations, the focal overtake remains unchanged, while progressively more surrounding vehicles are included. The 2-point case isolates the direct interaction between the two overtaking vehicles, the 5-point case embeds it within the immediate traffic context, and the 10-point case places it within a substantially richer interaction structure. The corresponding animations in the supplementary materials confirm that the same focal

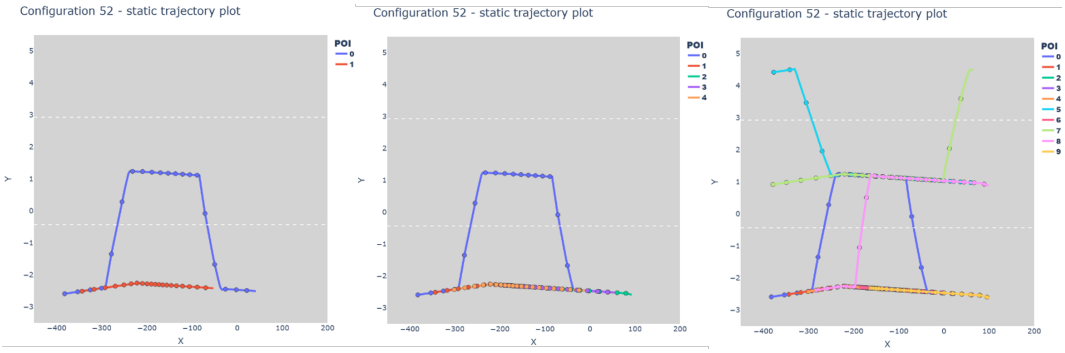


Figure 3 Configuration 52 represented at three levels of complexity (2-, 5-, and 10-point representations).

manoeuvre is preserved. This controlled design allows changes in the LLM output to be attributed primarily to representational complexity rather than to differences between manoeuvres.

3.2 From trajectory to qualitative representation

The increase in contextual detail is also visible in the qualitative encoding. Figure 4 shows the lateral-descriptor inequality matrices for configuration 52 at 2, 5, and 10 moving points. The matrices are visualised as heatmaps: green cells represent $<$, yellow cells $=$, and red cells $>$. Thick black lines mark point-specific row and column blocks, making pairwise relations easier to identify.

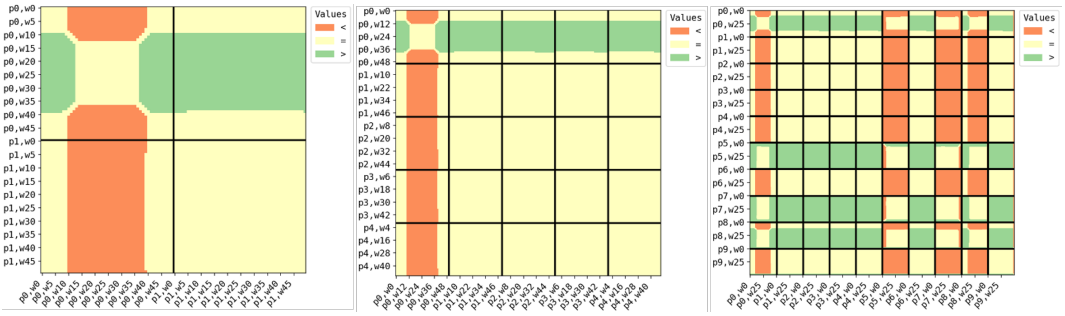


Figure 4 Heatmap representation of the lateral-descriptor inequality matrices for configuration 52 at 2, 5, and 10 moving points. Colours indicate qualitative relations ($<$, $=$, $>$), while the thick black lines mark the boundaries between point-specific row and column blocks.

In the 2-point representation, the matrix contains no surrounding traffic and captures only relations involving the two overtaking vehicles. The top-right submatrix shows how p_0 changes relative to p_1: both vehicles first occupy the same lane-level relation, then separate laterally during the overtaking phase, and finally return to the same lane-level relation. In the 5-point and 10-point representations, the same focal overtake is preserved but relations with surrounding vehicles are added. The representation therefore expands from a two-vehicle interaction to a richer local traffic configuration. In the 5-point representation, p_0 is the main changing actor, while the remaining vehicles form a comparatively stable surrounding group. In the 10-point representation, the colour pattern becomes more fragmented, especially for p_5, p_7, and p_8, making the overall structure harder to summarise at a glance.

This progression illustrates the central trade-off: with few points, the manoeuvre is readable but structurally limited; with more points, the encoding reflects a broader traffic scene but becomes harder to summarise. The inequality matrices therefore function as the key intermediate representation between raw trajectories and LLM interpretation.

3.3 LLM interpretation of the qualitative encoding

We focus on the merged manoeuvre interpretations, as these provide the most readable output of the staged API workflow. The longitudinal and lateral descriptor encodings were first interpreted separately, then combined into one final interpretation. The evaluated merged API outputs, including the corresponding JSON responses and expert evaluations, are provided in the supplementary materials.

The merged outputs were assessed against expert reference interpretations using the procedure described above. In the 2-point representation, the model produced a compact and strongly aligned overtaking narrative: p_0 starts behind p_1, overtakes it once, remains ahead, and shows a temporary lateral offset followed by realignment. This case was judged "Strong" on all four criteria by the experts.

In the 5-point representation, the output was richer and identified p_0 as the main changing actor within a comparatively stable surrounding group. It reconstructed a manoeuvre in which p_0 separates laterally, overtakes p_1 and later p_4, and then realigns. This case was judged "Strong" overall, with a minor reservation concerning spatial precision.

In the 10-point representation, the output produced the most nuanced structural reading. Rather than reducing the scene to a single overtaking label, it identified a stable lateral core and a smaller set of actors whose grouping and ordering changed over time. This case remained strong in completeness and interpretive adequacy, but was more limited in relational correctness and spatial precision than the lower-complexity cases.

Across the three levels, the clearest differences appeared in spatial precision and compactness. The 2-point representation yielded the clearest focal manoeuvre, the 5-point representation provided the best balance between clarity and contextual richness, and the 10-point representation supported the most nuanced but least concise structural reading. These findings are summarised in Table 1.

Table 1 Consolidated two-expert evaluation of merged LLM outputs across representation levels. Rel. corr. = relational correctness; Spatial prec. = spatial precision; Compl. = completeness; Interp. adequacy = interpretive adequacy.

Level	Rel. corr.	Spatial prec.	Compl.	Interp. adequacy	Overall judgement
2-point	Strong	Strong	Strong	Strong	Strongly aligned with expert interpretation
5-point	Strong	Adequate	Strong	Strong	Broadly aligned with minor issues
10-point	Adequate	Adequate	Strong	Strong	Broadly aligned with minor issues

4 Discussion and conclusion

This paper explored whether qualitative spatial encodings can support LLM-based interpretation of overtake manoeuvres. By representing the same overtaking event with 2, 5, and 10 moving points, the study isolates how increasing contextual complexity affects interpretation.

The staged API workflow further separates descriptor-level interpretation from the final merged manoeuvre summary, making the production of the natural-language output more transparent.

The results suggest that PDP-based inequality matrices provide a useful intermediate step between raw trajectories and natural-language description. The 2-point representation was easiest to interpret but offered limited context. The 5-point representation enriched the same manoeuvre with surrounding vehicles while remaining coherent, giving the best balance between relational richness and interpretability. The 10-point representation produced the most detailed structural reading, identifying a stable core and changing actors, but the description became less concise and less unified.

The contribution of the paper lies less in benchmarking LLM performance than in showing how the form of the encoding influences the interpretation produced. The LLM is not treated as a traffic-analysis model in its own right, but as a way to examine whether the symbolic structure captured by inequality matrices can be translated into coherent natural-language descriptions. In this setting, success means reconstructing the main relational structure of the manoeuvre in a spatially coherent and encoding-grounded way.

The study remains exploratory. Its main limitation is the use of a single illustrative configuration, which allows a controlled comparison across representation levels but does not yet show robustness across manoeuvres or interaction structures. The selected case was drawn from a broader collection of approximately 120 overtaking configurations per representation level. Extending the analysis to this larger set would support benchmark-style evaluation across cases, prompts, models, and representational variants. Future work could also extend the framework to merging, following, lane changes without passing, and other domains where moving point configurations are interpreted through changing spatial relations, such as sports or interaction-based movement analysis.

5 Data and materials availability

The supplementary materials are available via Zenodo: `https://doi.org/10.5281/zenodo.20729343`. They include the API code, representative request–response examples, evaluated merged outputs, expert evaluation materials, and derived qualitative encodings. Due to source restrictions, raw trajectory data are not shared.

References

1 C. Anghel, A. A. Anghel, E. Pecheanu, M. V. Craciun, A. Cocu, and C. Niculita. PEARL: A rubric-driven multi-metric framework for LLM evaluation. *Information*, 16(11), 2025. doi:10.3390/info16110926.

2 M. Berghaus, S. Lamberty, J. Ehlers, E. Kalló, and M. Oeser. Vehicle trajectory dataset from drone videos including off-ramp and congested traffic – analysis of data quality, traffic flow, and accident risk. *Communications in Transportation Research*, 4, 2024. doi:10.1016/j.commtr.2024.100133.

3 J. Chen, A. G. Cohn, D. Liu, S. Wang, J. Ouyang, and Q. Yu. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(1):106–136, 2015. doi:10.1017/S0269888913000350.

4 A. G. Cohn. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8, 2023. doi:10.48550/arXiv.2309.15577.

5 A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions. In *Leibniz International Proceedings in Informatics*, 2024. doi:10.4230/LIPIcs.COSIT.2024.28.

6 A. G. Cohn and J. Renz. Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*, pages 551–596. Elsevier, 2008. doi:10.1016/S1574-6526(07)03013-1.

7 Eirinaios-Odysseas Gardelakos, Vasileios Kyriakopoulos, Despina-Athanasia Pantazi, Orestis-Minas Kapopoulos, Maria Tsourma, and Manolis Koubarakis. Can large reasoning models reason about spatial relations? In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '25, pages 81–91, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3764912.3770841.

8 M. Hojati and R. Feick. Large language models: Testing their capabilities to understand and explain spatial concepts. In *Leibniz International Proceedings in Informatics*, volume 315 of *LIPIcs*, 2024. doi:10.4230/LIPIcs.COSIT.2024.31.

9 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark, 2024. doi:10.48550/arXiv.2401.03991.

10 A. Qayyum. *A Qualitative Representation of Moving Point Configurations with Illustrations in Micro Traffic Analysis*. PhD thesis, Ghent University, Faculty of Sciences, 2022. URL: http://hdl.handle.net/1854/LU-8750343.

11 A. Qayyum, B. De Baets, M. S. Baig, F. Witlox, G. De Tré, and N. Van de Weghe. The point-descriptor-precedence representation for point configurations and movements. *International Journal of Geographical Information Science*, 35(7):1374–1391, 2021. doi:10.1080/13658816.2020.1864378.

12 A. Qayyum, B. De Baets, L. De Cock, F. Witlox, G. De Tré, and N. Van de Weghe. Application of the point-descriptor-precedence representation for micro-scale traffic analysis at a non-signalized T-junction. *Geo-Spatial Information Science*, 26(3):406–430, 2023. doi:10.1080/10095020.2022.2069520.

13 T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, P. Mathur, G. E. Cacciamani, C. Sun, Y. Peng, and Y. Wang. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1), 2024. doi:10.1038/s41746-024-01258-7.

14 N. Van de Weghe. *Representing and Reasoning about Moving Objects: A Qualitative Approach*. PhD thesis, Ghent University, 2004.

15 N. Van de Weghe, A. G. Cohn, P. De Maeyer, and F. Witlox. Representing moving objects in computer-based expert systems: the overtake event example. *Expert Systems with Applications*, 29(4):977–983, 2005. doi:10.1016/j.eswa.2005.06.022.

16 N. Van de Weghe, A. G. Cohn, G. De Tré, and P. De Maeyer. A qualitative trajectory calculus as a basis for representing moving objects in geographical information systems. *Control and Cybernetics*, 35(1):97–119, 2006.

17 A. Yang, C. Fu, Q. Jia, W. Dong, M. Ma, H. Chen, F. Yang, and H. Wu. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, 39(9):2009–2044, 2025. doi:10.1080/13658816.2025.2490701.

18 Jiaxing Zhang, Wen Xiao, Benjamin Coifman, and Jon Mills. Vehicle tracking and speed estimation from roadside lidar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, September 2020. doi:10.1109/JSTARS.2020.3024921.

Limits of Spatial Reasoning Using LLMs

Varis Ithivatana
Independent researcher, MSc Business Analytics Graduate, University of Exeter, UK

Brandon Bennett
School of Computing, University of Leeds, UK

Abstract

This paper investigates the limits of spatial reasoning of LLMs in a grid-based warehouse environment, with ground-truth answers computed by simulation. We test two kinds of task: executing given movement instructions, and generating shortest action sequences for planning problems. In the execution tasks, an agent must follow egocentric and allocentric actions and report its final absolute direction relative to a target object. In the planning tasks, the model must find a shortest route, sometimes with object carrying, allocentric descriptions, or ordered pallet-pushing goals. We evaluate 100 instances per experiment using **GPT-5-mini**, and compare a smaller planning set with **ChatGPT-5.2** (extended thinking), **Gemini-Pro-3.1**, and **Grok-Expert**. The results show high accuracy on short and moderately long instruction-following tasks, but reduced accuracy on long trajectories with a much sharper drop on planning tasks. Static landmark labels in the form of coloured walls improve robustness, compared with problems described in terms of cardinal directions, while planning with ordered object-pushing remains difficult even for stronger models.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Spatial and physical reasoning

Keywords and phrases Spatial Reasoning, Large Language Models, Frame of Reference, Path Integration, Planning

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.17

Category Short Paper

Supplementary Material *Software (Source code and experiment results)*: `https://github.com/vi217/spatial-reasoning-llms`

Declaration on GenAI/LLM use Generative AI assistants were used during this work (2025–2026): Google **Gemini 2.5 Pro** and **Gemini 3.1 Pro** helped write a substantial part of the study's code (instance generation, the warehouse simulator and breadth-first-search ground-truth routines, and the scoring scripts), and Anthropic **Claude Opus 4.8** helped organise and refactor the code and prepare the code repository. All AI-generated code was read, executed, and tested by the authors, and the simulation and ground-truth logic was independently verified by a co-author. The authors retain full responsibility for the code, the results, and the content of this paper.

1 Introduction

Large language models (LLMs) perform well on many tasks that appear to require multi-stage inference, planning, or structured reasoning. However, it remains poorly understood what their underlying capabilities and limitations are, or how exactly such behaviour is achieved. The spatial domain provides useful test cases that involve combining pieces of information with complex interactions. Moreover, many spatial reasoning problems have clear ground-truth solutions that can be calculated by simulation or reasoning algorithms.

Our experiments use a warehouse domain with agents, pallets, racks, and forklifts occupying cells in a fixed grid. This setting is deliberately simple: it avoids visual ambiguity and allows actions to be interpreted as precise coordinate updates. At the same time, it supports several different kinds of spatial difficulty, including egocentric movement, allocentric

© Varis Ithivatana and Brandon Bennett;
licensed under Creative Commons License CC-BY 4.0

direction names, object movement, obstacle handling, object-relative orientation, and explicit planning. The main aim is not to introduce a new benchmark score, but to locate some points at which current LLM spatial reasoning behaviour begins to fail.

Benchmarks such as StepGame [5] and DecompSR [3] test multi-hop reasoning involving spatial relations and connect with broader work on compositional generalisation in neural models [2, 7]. In January 2024 Cohn and Blackwell [1] showed that models available at the time performed poorly on simple cardinal-direction questions. Current models are far more accurate on such short problems: in January 2026 we re-ran both of their datasets on **GPT-5-mini** and obtained 95% on the original small problem set and 88% on the larger version of the same problem set. We therefore examine longer, dynamic tasks where a model must reason about a spatial state that is updated by many actions.

2 Methodology

Our experiments concern a warehouse situation represented by a 10×10 grid, with cells indexed by x, y-coordinates. Each test specifies an initial agent position and facing, a target object, and a sequence of actions. The basic egocentric action set is $\mathbf{Ego} = \{F, B, SL, SR, TL, TR\}$, representing forward, backward, shift-left, shift-right, turn-left and turn-right. Some experiments also use diagonal moves and object-relative facing commands. For instruction-following tasks, the required output is one of $\mathbf{Card} = \{N, NE, E, SE, S, SW, W, NW\}$, expressing the direction from the agent to a named object in the final situation. Correct answers for instruction-following tasks are computed by a simulator implemented in Python. Ground-truth shortest paths for planning tasks are computed by breadth-first search. We used the same 10×10 warehouse layout for all experiments to hold the environment constant, so that performance differences reflect changes in the action rules and task requirements rather than the underlying map or grid size; varying the grid size and setting is left to future work.

2.1 Experiments

The evaluation consists of five experiments of increasing complexity. **Exp-1** is a baseline path-integration task where an agent executes a sequence of egocentric rotations and translations and reports the cardinal direction of a landmark relative to its final position. **Exp-1.1** replaces cardinal direction names by colour labels attached to the warehouse walls. A further variant, **Exp-1.2**, replaces only the cardinal direction names with coloured walls but keeps the egocentric diagonal moves; it is included for token-usage comparison at $k = 120$ only.

Exp-2 adds object movement and drop events: a robot pushes an object, a product is dropped after a specified unit push, and later agents transport it. **Exp-3** adds blocked "spill" cells and a detour rule. If a move would enter a blocked cell, it is replaced by a specified detour movement. The detour specification produces a sensible avoidance only when the worker approaches the spill from the West or East; when travelling North or South into a spill, the detour leaves the worker either one cell back from where it started or still adjacent to the obstacle. **Exp-3** therefore tests comprehension of a particular avoidance instruction rather than a general capacity to reason about avoidance and detours. **Exp-4** inherits the spill rule from **Exp-3** and also replaces some egocentric turning with object-relative orientation, so the agent must turn to face the current location of a specified object.

Whereas **Exp**s **1**–**4** mainly test state update under a supplied action sequence, **Exp**s **5**, **5.1**, **5.2** and **5.3** test the ability to construct an action sequence. This separation is useful because planning failures can otherwise be mistaken for failures of spatial interpretation. For **Exp-5** the model must output a shortest valid action sequence from a start position to a

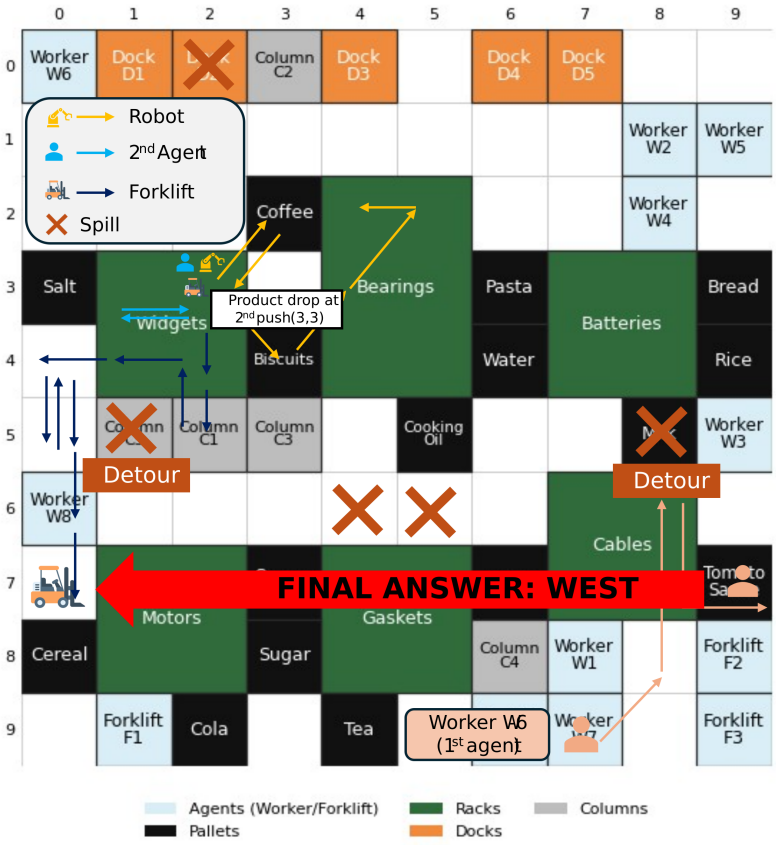


Figure 1 Warehouse grid layout, entity locations and example movement sequences used in the experiments. Crosses indicate locations of "spill" events.

target location. **Exp-5.1** adds object carrying; **Exp-5.2** adds allocentric "riddles" involving rack-facing relative to another agent; and **Exp-5.3** adds ordered pallet-pushing goals, where a push is valid only under specified hand, facing, and occupancy conditions.

2.2 Dataset, models and prompts

For each experiment, **GPT-5-mini** was evaluated on 100 instances, using two random seeds with 50 instances from each. For **Exps 1–4**, action sequences were generated from random-walk controls with a minimum length of five actions. The action distribution for the random walk was set to: turn 0.45, lateral shift 0.25, diagonal move 0.25, with the residual mass on forward/backward. These values were chosen so that all six action types appear within most instances and the agent's facing direction (the key state quantity the model must track) changes frequently. We did not perform a sensitivity analysis on these proportions and report them so that any reader can reproduce or vary them. The seeds 7 and 8 are arbitrary fixed values used for reproducibility. For the planning variants, we additionally evaluated **ChatGPT-5.2** (extended thinking), **Gemini-Pro-3.1**, and **Grok-Expert** on 30 instances per condition. These additional runs were performed through web interfaces and were therefore smaller, but they are useful for distinguishing limitations of **GPT-5-mini** from limitations that persist in stronger models.

Table 1 Summary of the experimental conditions. Unless otherwise stated, **GPT-5-mini** was evaluated on $n = 100$ instances and stronger models on $n = 30$ instances per experiment.

Experiment	Main additional requirement	Task type
1	Egocentric grid tracking, $k = 5$–300	execute
1.1	Cardinal directions replaced by coloured walls	execute
1.2	Coloured walls, keeping egocentric diagonals, $k = 120$	execute
2	Object pushed, dropped, picked up and transported	execute
3	Blocked spill cells with detour rule	execute
4	Object-relative orientation commands	execute
5	Shortest route to target	plan
5.1	Pick up, deliver, and return	plan
5.2	Allocentric rack-facing descriptions	plan
5.3	Ordered pallet-pushing goals	plan

All prompts are zero-shot. We avoided few-shot tests because in-context demonstrations risk leaking the structure of a valid trace (e.g. the JSON schema and canonical detour pattern) into the model response, converting **Exp-3** from a comprehension test into a copy-and-adapt task. We acknowledge that few-shot prompting may improve absolute accuracy and leave a controlled comparison for future work. The system prompt gives the coordinate convention, action rules, entity positions and output format. User prompts give the instance-specific start state, target, movement sequence and any additional constraints. For **Exp**s **2**–**4**, we also requested trace output to inspect error sources. Scoring, however, was based only on the final direction or plan, not the trace. The traces were purely diagnostic, allowing us to distinguish, for example, a wrong final direction caused by an early movement update from one caused by a later drop-location error.

2.3 Statistical procedure

Instances are generated independently for each experiment and condition, so all reported comparisons are between independent samples; we use two-proportion z-tests for differences and report 95% Wilson intervals per proportion. For **GPT-5-mini**, each experiment pools seeds 7 and 8 ($n = 50$ each, $n = 100$ total). Multiple comparisons among the stronger models (Exp 5–5.3) use the Holm correction.

2.4 Reproducibility considerations

Commercial LLMs may receive silent updates between versions, so our numerical results are just a snapshot of February 2026 behaviour. **GPT-5-mini** API runs (16–25 Feb. 2026) used the OpenAI `/responses` endpoint with model name `GPT-5-mini`, temperature 1.0, API `seed` 42, and top-p at the default. Every prompt and response is stored with the model name, prompt text, generation seed, token usage and elapsed time. Web-interface evaluations of **ChatGPT-5.2** (extended thinking), **Gemini-Pro-3.1** and **Grok-Expert** (20 Feb. 2026) are recorded as screenshots and transcripts; these are inherently less reproducible than API runs, so the more nuanced quantitative claims rely only on the API-based **GPT-5-mini** results. We plan to extend evaluation to open-weight models so the pipeline can be re-run independently of any commercial provider.

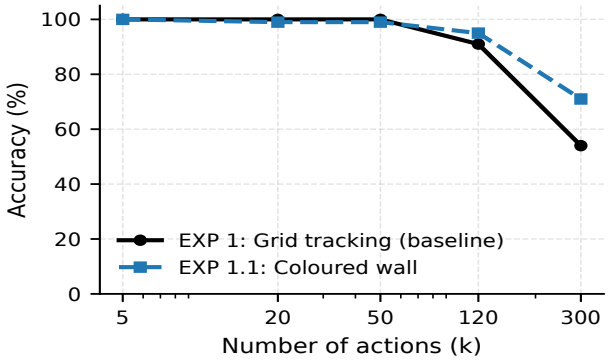


Figure 2 Accuracy with trajectory length k for **Exp-1** (baseline) and **Exp 1.1** (coloured walls).

3 Results

3.1 Instruction following

On the baseline trajectory task, **GPT-5-mini** remains highly accurate for short and medium action sequences. Accuracy is between 0.90 and 1.00 for $k = 5$–120, but falls at $k = 300$. The coloured-wall variant improves the longest-sequence result: at $k = 300$, baseline accuracy is 0.54, while coloured walls is 0.71 (Figure 2). The difference of 0.17 is statistically significant, with a 95% confidence interval of $[0.04, 0.30]$. Coloured walls also reduce median token usage relative to the baseline at $k = 120$ and above, indicating that the explicit landmark cue improves accuracy and reduces computation. **Exp-1.2**, with cardinal directions and egocentric diagonals, uses more tokens than **Exp-1.1** at $k = 120$, suggesting that replacing the diagonal moves contributes a substantial part of the token saving.

Pooled across both seeds ($n = 100$), accuracy is 0.99 in **Exp-2**, 0.88 in **Exp-3**, and 0.80 in **Exp-4**. The drop from **Exp-2** to **Exp-3** is statistically significant ($[0.04, 0.19]$). The further drop from **Exp-3** to **Exp-4** is not statistically significant at the 95% level ($[-0.02, 0.18]$, which includes zero).

Trace error analysis

For **Exp-3** and **Exp-4** we ran the trace prompt to identify where errors occur, using the seed-7 set ($n = 50$ per experiment). Errors in a small set of samples ($n = 5$ and $n = 11$ respectively) were categorised by a single annotator. Hence, we present the analysis as qualitative and illustrative rather than as a quantitative claim. To keep the categories mutually exclusive we use the following operational rules. "Incorrect movement action" applies when the trace's next coordinate disagrees with the simulator's and the step is not preceded by a spill cell; "failure to apply detour rule" applies when the trace enters a spill cell or omits the prescribed four-step detour. "Incorrect product-drop location update" applies when the recorded drop position differs from the simulator's at the specified step; "product dropped at incorrect step k" applies when the drop occurs at any step other than k. The four categories partition the failure space, so each trace step is assigned to exactly one. Table 2 shows the breakdown.

In Experiment 3, none of the five errors arose from misapplying the spill or detour rule. They came from updating the wrong agent coordinate, or tracking the product-drop position incorrectly. We interpret the Experiment 3 drop as a state-tracking failure under a longer

Table 2 Error type breakdown from full-trace analysis for Experiments 3 and 4 (seed 7, $n = 50$). Single annotator; qualitative, illustrative.

Experiment	Error type	Count	Share
Exp-3	Incorrect movement actions	2	40%
	Incorrect product-drop location update (used original forklift position)	2	40%
	Product dropped at incorrect step	1	20%
Exp-4	Incorrect movement actions (not including cases due to blockage)	7	64%
	Failure to detect blockage / apply detour rule	3	27%
	Incorrect product-drop location update	1	9%

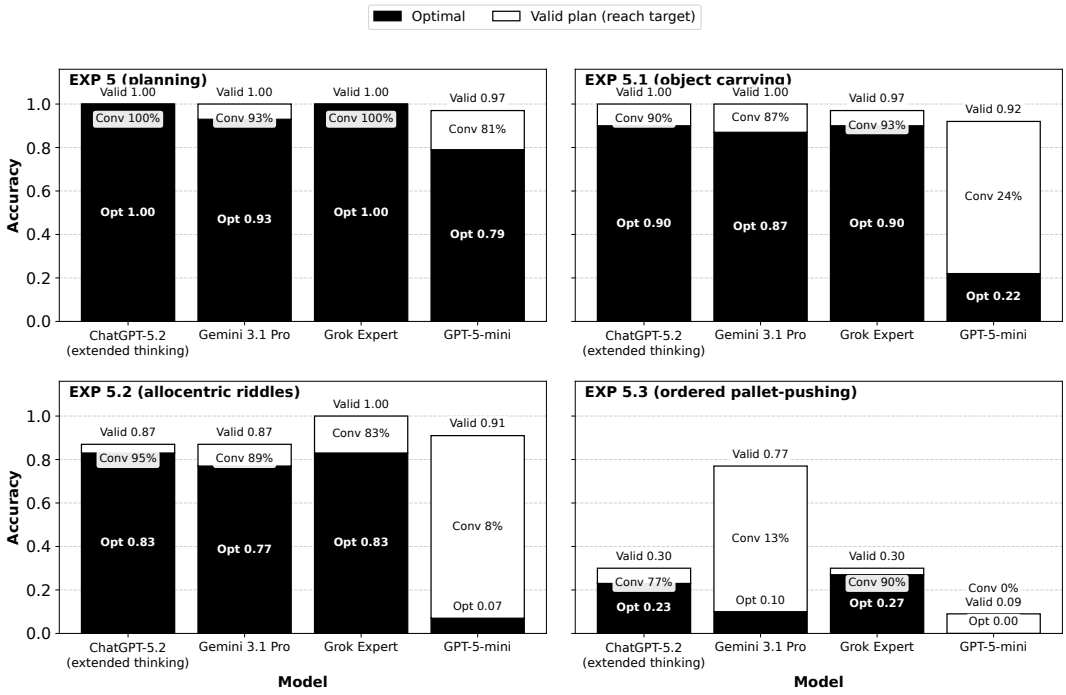


Figure 3 Planning accuracy for Experiments 5–5.3 by model. Black bars show optimal; outlined bars show valid plan minus optimal. Conversion rate = optimal / valid plan.

and more rule-laden prompt, rather than a rule-comprehension failure. In Experiment 4 the dominant pattern is incorrect movement actions (7 of 11), which suggests a separate failure mode in vector computation under variable binding.

3.2 Planning

Planning is substantially more difficult. **Exp-5** answers are scored as correct only when they reach the target by a shortest route. **GPT-5-mini** achieves 0.79 on the basic planning task, but falls to 0.22 with object carrying, drops further to 0.07 with allocentric riddles, and reaches 0.00 in the ordered pallet-pushing task (Figure 3). Its valid-plan rate is often higher than its optimal-plan rate, indicating that the model can frequently reach the target but fails to find a shortest route. The fraction of its valid plans that are also optimal, the *conversion rate* ("Conv" in Figure 3), is therefore low.

The stronger models perform much better on the first three planning variants. In **Exp-5** **ChatGPT-5.2** (extended thinking) and **Grok-Expert** both achieve optimal-route rate 1.00, and **Gemini-Pro-3.1** reaches 0.93. On **Exp-5.2**, **ChatGPT-5.2** achieves optimal-route rate 0.83,

Grok-Expert 0.83, and **Gemini-Pro-3.1** 0.77. All three nevertheless show a sharp decline on **Exp-5.3** in optimal-route rate: **ChatGPT-5.2** drops from 0.83 to 0.23 (a difference of 0.60), **Grok-Expert** from 0.83 to 0.27 (0.56), and **Gemini-Pro-3.1** from 0.77 to 0.10 (0.67), with all three differences statistically significant. A model-specific pattern is also worth noting: for **Exp**s **5.1**–**5.3**, **Gemini-Pro-3.1** consistently shows the lowest optimal-route rate among the stronger models while maintaining a high valid-plan rate, particularly on **Exp-5.3**, indicating that it reliably reaches the target without finding the shortest route. Therefore, under the specific prompting, representation, and models tested, ordered object pushing marks a limitation of current spatial planning rather than a general bound; because the collapse appears across three independent frontier models, we expect it to hold for newer ones, though confirming this is left to future work.

A qualitative pattern in the **GPT-5-mini** planning errors is avoidance of diagonal moves: the model produces a valid route composed mainly of cardinal moves, although a shorter route using diagonals exists. This implies a non-optimal route-generation strategy rather than systematic search over the action space. The optimality requirement is therefore crucial: without it, **Exp-5** would be much easier.

4 Discussion

The results suggest that generating an action sequence to achieve a goal is much more difficult for LLMs than inferring spatial relations holding after executing a specified action sequence. **GPT-5-mini** is often robust when following long sequences, even when several kinds of spatial relations and events are involved. Its limitations emerge for very long trajectories, but much more acutely for planning, where a shortest route must be found rather than merely followed. The coloured-wall result is notable: replacing cardinal names by arbitrary colour labels improves performance on very long trajectories. We had expected better performance with the more conventional way of expressing directions. Perhaps the colour coding forces the LLM into a more algorithmic mode, rather than relying on statistical correlations that may be insufficient for handling long trains of reasoning.

The planning failures are consistent with earlier observations that LLMs do not reliably implement exhaustive search [6]. A valid but non-optimal route can be generated by following a locally reasonable pattern, whereas an optimal route requires comparing alternatives. The ordered pallet-pushing task adds further constraints: the model must interleave route choice, object status, facing direction, and ordering. This combination sharply reduces performance even for stronger models. **Exp-5.3** therefore marks a limitation of current spatial planning under the specific prompting, representation, and model set tested. We note one design limitation: the later experiments add planning on top of spatial reasoning, so a failure in **Exp-5.3** is more accurately a failure of planning under spatial and object-state constraints than a pure spatial-reasoning failure. This remains relevant in practice, since many real uses of spatial understanding involve considering alternative possibilities, but the distinction matters when interpreting the results.

5 Conclusion

We evaluated LLM spatial reasoning in a grid environment with simulation-derived ground truth. **GPT-5-mini** performs well on many instruction-following tasks, but accuracy declines on very long trajectories and when dynamic object-relative orientation is added. Compared to instructions given in terms of cardinal directions, static landmarks in the form of coloured walls

improved long-sequence robustness and reduced token use at $k = 120$ and above. Planning is harder: valid routes are often produced, but shortest routes missed. Ordered pallet-pushing goals cause a sharp drop for all models tested. This is not surprising, since spatial action planning is intrinsically difficult. However, the observation that once a certain complexity threshold is reached, reasoning accuracy drops very sharply (without any indication in the actual output) could be a serious issue for any system relying on an LLM for spatial planning.

Future work should more clearly separate different aspects of spatial reasoning, as well as investigating their combinations in a systematic way; in particular, varying the grid size and setting (e.g. larger or outdoor layouts) would help disentangle sequence-length effects from grid-size effects. Going beyond simple grid-based situations to richer spatial puzzle domains [4] will further illuminate the spatial reasoning capabilities and limitations of LLMs, as will evaluating newer model snapshots (e.g. Claude Opus and post-GPT-5.4 models) to test whether the failure modes we identify persist as models improve.

References

1 Anthony G Cohn and Robert E Blackwell. Evaluating the ability of large language models to reason about cardinal directions (short paper). In *Proceedings of the 16th conference on Spatial Information Theory (COSIT)*, volume 315. Schloss Dagstuhl, LIPIcs, 2024. doi:10.4230/LIPIcs.COSIT.2024.28.

2 Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020. doi:10.1613/JAIR.1.11674.

3 Lachlan McPheat, Navdeep Kaur, Robert Blackwell, Alessandra Russo, Anthony G. Cohn, and Pranava Madhyastha. Decompsr: A dataset for decomposed analyses of compositional multihop spatial reasoning, 2025. doi:10.48550/arXiv.2511.02627.

4 Paulo E Santos, Pedro Cabalar, Zoe Falomir, and Thora Tenbrink. Representing and solving spatial problems. *Spatial Cognition & Computation*, 25(1):1–14, 2025. doi:10.1080/13875868.2024.2411052.

5 Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. Pre-print: arXiv:2204.08292. doi:10.48550/arXiv.2204.08292.

6 Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models: A critical investigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

7 Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability. In *First Conference on Language Modeling*, 2024.

A Data, Code, and Further Results Availability

Our code repository including prompt sets, raw model outputs, and an extensive technical report are available from `https://github.com/vi217/spatial-reasoning-llms`.

Our simulator scripts enable the results to be reproduced as faithfully as possible while the underlying model version remains available. The same tests could also be run on future models.

A Cross-Linguistic Evaluation of Cardinal Direction Reasoning in LLMs

Panagiota Gyftou
Dept. of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

Robert E. Blackwell
School of Computer Science, University of Leeds, UK

Manolis Koubarakis
Dept. of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

Anthony G. Cohn
School of Computer Science, University of Leeds, UK
Alan Turing Institute, London, UK

Abstract

We present a cross-linguistic evaluation of cardinal direction reasoning in large language models. We extend an earlier cardinal direction benchmark, translating the questions into Greek and French and compare 10 LLMs on the new benchmarks.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning

Keywords and phrases Large Language Models, Spatial Reasoning, Cardinal Directions

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.18

Category Short Paper

Supplementary Material *Dataset (Dataset)*: `https://github.com/panagiotagyft/cosit2026`
archived at `swh:1:dir:9ef27812d94d1315e5c45dc3ae396cde58141d0a`

Funding This work was partially supported by the Marie Skłodowska-Curie Actions Staff Exchanges fund (project ThirdWave, GA no. 101236394), the Fundamental Research priority area of The Alan Turing Institute, the Economic and Social Research Council (ESRC) under grant ES/W003473/1, and the EPSRC under grant EP/Z003512/1.

Acknowledgements This work was supported by computational time granted from the National Infrastructure for Research and Technology S.A. (GRNET S.A.) in the National HPC facility – ARIS – under project ID pa260202-A Cross-Linguistic Evaluation of Cardinal Direction Reasoning in LLMs.

1 Introduction

Many claims have been made about the reasoning capabilities of Large Language Models (LLMs) [10] and many benchmarks have been developed to evaluate LLM reasoning performance e.g., Grade School Math 8K (GSM8K) [11]. The majority of widely used LLM evaluation benchmarks remain English-centric, with multilingual benchmarks representing a minority (roughly 20-30%) of well-known benchmarks (as an example, MGSM [8] extends GSM8K to 10 typologically diverse languages). As a result, the reasoning capabilities of LLMs in other languages remain comparatively underexplored. The reasoning problem on which we concentrate in this paper is *spatial reasoning about cardinal directions (CDs)* [1, 2]. CDs are used to describe how regions of space are placed relative to one another (e.g., region

© Panagiota Gyftou, Robert E Blackwell, Manolis Koubarakis, and Anthony G Cohn;
licensed under Creative Commons License CC-BY 4.0

a is north of region b). They are an important case of qualitative spatial relations which are used to model space similarly to the way humans do. They have been found to be useful in GIS, autonomous navigation and robotics and computer vision.

In order to explore the reasoning capabilities of modern LLMs in languages other than English, this paper concentrates on two European languages: French and (Modern) Greek. French is a high-resource language but not as dominant as English. As an example, there are only 4.6052% documents with primary language French in the Common Crawl latest version CC-MAIN-2026-12[1], which is one of the large corpora used to train LLMs, compared to 41.0588% English documents. Greek, on the other hand, is substantially under-represented (0.5494% of the total number of documents).

Despite being major European languages, French and Greek present unique hurdles for LLMs: (i) morphological complexity: French and Greek are highly inflected (words change form based on case, gender and number). This requires more data for a model to learn the rules of the language compared to English. (ii) script heterogeneity (in the case of Greek): Because Greek uses its own alphabet, it does not benefit from lexical sharing the way French does (e.g., a model learns the word "information" in English and instantly understands "information" in French). (iii) tokenization inefficiency: many state-of-the-art LLMs break French and Greek words into more tokens than English words because they use tokenizers optimized for English [3, 15, 6]. This makes processing French and Greek more expensive and slightly slower for the model.

The contributions of this paper are the following:

1. We develop the first two benchmarks for reasoning about CDs in French and Greek by translating the (English) benchmark of [1, 2] into these two languages. The new benchmarks consist of 7680 questions in each language.
2. We evaluate the reasoning capabilities of 10 LLMs on the new benchmarks, in order to explore how language affects spatial reasoning performance in state-of-the-art LLMs. We have selected a variety of models: one commercial multilingual LLM with strong reasoning capabilities (*GPT-5.2*), three smaller multilingual LLMs with enhanced reasoning capabilities but not a dedicated reasoning architecture (*Kimi-k2-0905*, *Deepseek V3.2* and *Llama-3.3-70B*), the European multilingual models *EuroLLM-22B* and *EuroLLM-9B*, the state of the art models for the Greek language *Meltemi-7B* and *Krikri-8B*, the model *Lucie-7B* for French, and the small multilingual model *Mistral Nemo*. Our experiments show that, in general, spatial reasoning in English yields better accuracy for the majority of examined LLMs, given that they have been trained mostly on English corpora. French and Greek showed consistently strong performance, even outperforming English in the *GPT-5.2* reasoning model. Generally, the models struggle with intercardinal directions, suggesting that compound words increase reasoning complexity. Additionally, sentence clarity plays a crucial role in the quality of responses, as complexity increases, accuracy tends to decrease. Finally, variations in locomotion type and grammatical person do not seem to affect the responses.

2 Related work

Cohn and Blackwell tested LLM reasoning using a CD benchmark [1, 2] showing evidence of systematic errors and bias in LLM reasoning. The benchmark consists of 5760 questions and answers generated from a set of templates, testing an LLM's ability to determine the correct CD given a particular scenario (for more details see Section 3). However, they only

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

test English language questions and they do not test European LLMs, except for Mistral. It remains unclear whether their findings generalize across languages with different scripts, morphologies, and spatial expressions. Another recent work which considers reasoning about CDs is [7] where the authors evaluate 7 large reasoning models on the task of computing the transitivity table of the CD calculus of [19] which deals with regions and allows relations North, South, East, West and their combinations (e.g., Northwest), but only in English.

The question of whether language influences reasoning in LLMs has been studied by other researchers too. [5] analyzes the internal workings of Qwen1.5-1.8B-Chat in cross-lingual causal reasoning from both syntactic and semantic perspectives. [9] introduce the cross-lingual-thought prompting method to systematically improve the multilingual capability of LLMs. [13] introduced a novel methodology, an open-source framework for generating correct-by-construction natural language spatial reasoning tasks that enable decomposed evaluation of compositional abilities beyond mere accuracy, and DecompSR, a large-scale, customisable benchmark dataset (over 5 Million samples) for multi-hop compositional spatial reasoning. They also evaluated the performance of contemporary LLMs demonstrating that while LLMs show some linguistic resilience, they largely struggle with systematic and productive generalisation in spatial tasks and exhibit overgeneralisation tendencies. Finally, [4] introduces the multilingual dataset xSTREET that covers four tasks across six languages. The authors show that xSTREET demonstrates a gap in base LLM performance between English and non-English reasoning tasks. Then, they propose two methods (one for training time and one for inference time) to remedy this gap, building on the idea that LLMs trained on code are better reasoners.

In Europe, the most important language technologies activity at the moment is the Alliance for Language Technologies (ALT-EDIC). ALT-EDIC is a "European Digital Infrastructure Consortium supporting the development of language technologies[2]. It promotes excellence in language technologies and contributes to the preservation of Europe's linguistic diversity and cultural richness." In addition, it aims to strengthen Europe's technological and economic sovereignty and competitiveness, through the development of transparent and compliant open-source multilingual models. These models are tailored to all official European languages, as well as several languages of strategic international importance. In the context of ALT-EDIC, OpenEuroLLM is a family of LLMs developed in collaboration with the High Performance Language Technologies initiative[3] which has produced 38 monolingual models. These cover all European languages excluding Irish and Maltese.

Another important activity in Europe has been the development of EuroLLM[4] supporting all 24 official EU languages. A significant milestone is the collaboration between OpenEuroLLM and EuroLLM on MultiSynt, an open-source multilingual synthetic dataset for LLM pre-training. As the first effort of its kind, democratizes access to high quality training data, particularly for languages that remain underrepresented in the current LLM space[5].

For the French and Greek languages considered in this paper, there is also considerable recent LLM development activity. The most important LLMs with an emphasis on the French language are Lucie [15] and Cedille [18]. Lucie-7B is trained on equal amounts of data in French and English (roughly 33% each). Cedille is trained only on French data. We also note that the French company Mistral AI is a leading European player in the LLM space, with models that are multilingual rather than exclusively focused on French.

[2] https://www.alt-edic.eu/
[3] https://huggingface.co/collections/HPLT/hplt-20-monolingual-reference-models
[4] https://eurollm.io/
[5] https://openeurollm.eu/blog/multisynt-synthetic-training-data

Recent Greek LLMs include Meltemi-7B [12] and Llama-Krikri-8B [6]. The latter model is currently the state of the art for the Greek language. It has been trained on a corpus which includes 56.7 billion monolingual Greek tokens (62.3%), 21 billion monolingual English tokens (23.1%), 5.5 billion parallel data tokens (6.0%), and 7.8 billion math and code tokens (8.6%). It demonstrates substantial improvements for Greek (+10.8%) compared to Llama-3.1-8B [6]. Moreover, Llama-Krikri-8B-Base surpasses Meltemi-7B-v1.5 with a notable +11.6% average improvement across all benchmarks tested.

3 Experimental design

The benchmark questions from [1, 2] are generated from six templates each modulated with ten forms of locomotion (cycling, driving, hiking, jogging, perambulating, racing, riding, running, unicycling and walking), eight directions (north, south, east, west, north-east, north-west, south-east, south-west) and each of the six English person forms: first-person (I am), first-person plural (We are), second-person (You are), third-person singular (He is and She is), and third-person plural (They are). This gives us 6 templates $\times$ 10 locomotion forms $\times$ 6 person forms $\times$ 8 directions $\times$ 2 direction-variations = 5760 questions.

We take these templates and translate them into both Greek and French, which both have two more person forms. Moreover, verbs have to agree in person and number, and both Greek and French exhibit a richer inflectional system for person and number compared to English, where these differences are implicit. Specifically, English uses the single pronoun "You" for both the second person singular and plural. In contrast, both Greek and French employ distinct forms, "Εσύ" (Greek) and "Tu" (French), for the singular, and "Εσείς" and "Vous" for the plural. Additionally, the second person in both languages differs in formal and informal contexts. Specifically, the second person singular is used informally for people one knows or for children, whereas the second person plural can be used in the singular to express politeness. Similarly, although all three languages distinguish gender in the third person singular ("He"/"She", "Αυτός"/"Αυτή", "Il"/"Elle"), in the plural, English uses only "They", whereas Greek and French distinguish by gender: "Αυτοί"/"Ils" are used for masculine or mixed groups, and "'Αυτές'/"Elles" are used for groups of women. So our benchmarks for French and Greek consist of 7680 questions ($6 \times 10 \times 8 \times 8 \times 2$).

We adopt the same experimental setup as [1, 2] setting temperature to 0 with a fixed random seed (where available). We perform three runs per prompt to verify the consistency of the results across iterations. Each prompt is presented as a new LLM conversation so that models do not retain context from earlier prompts.

Table 1 Example question templates for each language.

English	T1	You are walking *[south]* along the *[east]* shore of a lake; in which direction is the lake?
	T2	You are walking *[south]* along the *[east]* shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake?
	T3	You are walking *[south]* along the middle of the *[east]* side of a park; in which direction is the bandstand located in the centre of the park?
	T4	You are walking *[east]* along the *[south]* side of a road which runs *[east to west]*. In which direction is the road?
	T5	You are walking *[south]* along the *[east]* shore of the island. In which direction is the sea?
	T6	You are walking *[south]* along the *[east]* shore of an island and then turn around to head back in the direction you came from, in which direction is the sea?
French	T1	Tu marches *[au sud]* le long de la rive *[est]* d'un lac. Dans quelle direction se trouve le lac?
	T2	Tu marches *[au sud]* le long de la rive *[est]* d'un lac et puis tu retournes pour revenir dans la direction d'où tu es venu(e). Dans quelle direction se trouve le lac?
	T3	Tu marches *[au sud]* au milieu du côté *[est]* d'un parc. Dans quelle direction se trouve la banderole située au centre du parc?
	T4	Tu marches *[à l'est]* le long du côté *[sud]* d'une route qui s'étend de *[l'est à l'ouest]*. Dans quelle direction se trouve la route?
	T5	Tu marches *[au sud]* le long de la côte *[est]* d'une île. Dans quelle direction se trouve la mer?
	T6	Tu marches *[au sud]* le long de la côte *[est]* d'une île et puis tu retournes pour revenir dans la direction d'où tu es venu(e). Dans quelle direction se trouve la mer?
Greek	T1	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* όχθης μιας λίμνης. Προς ποια κατεύθυνση βρίσκεται η λίμνη;
	T2	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* όχθης μιας λίμνης και στη συνέχεια γυρίζεις για να επιστρέψεις προς την κατεύθυνση από την οποία ήρθες. Προς ποια κατεύθυνση βρίσκεται η λίμνη;
	T3	Περπατάς *[νότια]* στη μέση της *[ανατολικής]* πλευράς ενός πάρκου. Προς ποια κατεύθυνση βρίσκεται η εξέδρα της μπάντας στο κέντρο του πάρκου;
	T4	Περπατάς *[ανατολικά]* κατά μήκος της *[νότιας]* πλευράς ενός δρόμου που εκτείνεται από τα *[ανατολικά προς τα δυτικά]*. Προς ποια κατεύθυνση βρίσκεται ο δρόμος;
	T5	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* ακτής ενός νησιού. Προς ποια κατεύθυνση βρίσκεται η θάλασσα;
	T6	Περπατάς *[νότια]* κατά μήκος της *[ανατολικής]* ακτής ενός νησιού και στη συνέχεια γυρίζεις για να επιστρέψεις προς την κατεύθυνση από την οποία ήρθες. Προς ποια κατεύθυνση βρίσκεται η θάλασσα;

4 Results

4.1 Accuracy of models across all languages

To compute accuracy we adopt the formula [17] which calculates the mean accuracy for a model while simultaneously determining a prediction interval to quantify accuracy variance across repeated evaluations. Therefore, in the heatmaps appearing in this section, we present accuracy values as a prediction interval (i.e., mean value ± a margin of error).

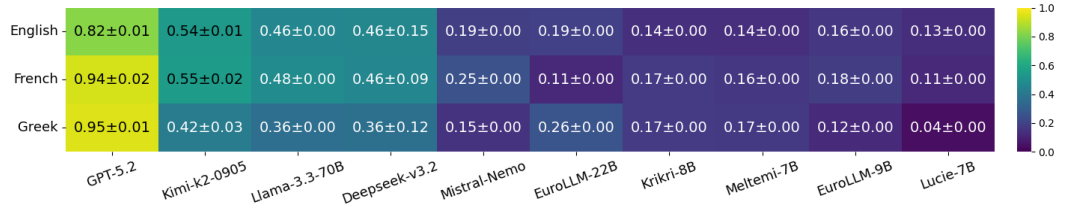


Figure 1 Comparison of model performance across three languages.

As we see in Figure 1, the best performing model across all languages is *GPT-5.2*, showing the highest accuracy in Greek (95%), closely followed by French (94%) and finally English (82%). It is surprising that the English language performance of GPT-5.2 reported here (0.82) is worse than that of o1 (0.92) reported by Cohn and Blackwell [2], despite GPT-5.2 being a newer OpenAI model. This discrepancy is particularly pronounced for template T4 questions: GPT-5.2 achieves a score of only 0.08, compared to 0.55 for o1. The superiority of French and Greek might be considered a surprising and counterintuitive result, given that *GPT-5.2* has been trained predominantly on English corpora. We conjecture that this result may be due to one or more of the following factors: (i) the fact that the translated questions may be less ambiguous compared to the English ones, e.g., with respect to person form, given that French and Greek are highly inflected languages. (ii) data contamination and overfitting given that the English dataset has been published on the Web since 2024 and it is may be part of the training data of *GPT-5.2* which has a knowledge cutoff of late 2025 (though the answers are in a separate file); in contrast, the French and Greek datasets are new so the model has to reason about them from first principles (unless the model were to translate into English for the reasoning), (iii) denoising: English training data is noisy in comparison with Greek and French which might be of higher quality (e.g., from Wikipedia) in a higher percentage, and (iv) more tokens per cardinal direction for French and Greek which could make processing less efficient but enable more detailed reasoning steps. (v) *GPT-5.2* uses a dynamic reasoning approach – i.e. by default, it auto-decides how much reasoning effort to devote to a problem; it might be that the presence of a language other than English makes it more likely to trigger a higher reasoning effort which may in turn lead to more accurate answers. Regarding the English dataset, excluding *GPT-5.2*, the leading models are *Kimi-k2-0905* with an accuracy of 54%, *Deepseek V3.2* with 46% and *Llama-3.3-70B* with the same accuracy of 46%. The reason for this lower performance is that these models have significantly fewer parameters than *GPT-5.2* and are not truly reasoning models with a specialized reasoning architecture; however, according to their developers, they have enhanced reasoning capabilities as opposed to pure LLMs. The leading models for Greek follow the same hierarchy as in English but with lower accuracy scores. *Kimi-k2-0905* achieved 42%, while *Deepseek V3.2* and *Llama-3.3-70B* scored 36%. On the other hand, for French, the models behave similarly to the English dataset: *Kimi-k2-0905* achieved

55% accuracy, while *Llama-3.3-70B* reached 48% and *Deepseek V3.2* 46%. It is notable that *Llama-3.3-70B* has fewer parameters than *Deepseek V3.2* yet it achieved the same or sometimes better accuracy on the French dataset.

The smaller models in Figure 1 exhibit different behaviors across the three languages. Specifically, for the European instruct models, we see that *EuroLLM-9B* performed better on the French dataset with 18% accuracy, compared to 16% and 12% for the English and Greek datasets, respectively (chance is 12.5%). In addition, for *EuroLLM-22B*, we observed better performance in the Greek language with 26%, while English showed a lower accuracy score of 19% and French 11%. *EuroLLM-9B* was trained initially on 50% English, with this percentage being reduced gradually during the next two phases of its training. The French dataset accounted for $\sim 5-6\%$ across all three phases, whereas Greek was only $\sim 0.5-1\%$; this explains these results [16]. *EuroLLM-22B* was also trained with 60% English, $\sim$9% French and $\sim$1% Greek, but in the last phases the authors of the EuroLLM models used the EuroFilter to choose the best quality data [14]. We conjecture that in the case of *EuroLLM-22B*, the Greek data was of better quality than the French. Furthermore, regarding the Greek models, *Meltemi-7B* and *Krikri-8B*, as anticipated, both showed superior performance in the Greek dataset, each reaching 17% accuracy. On the English dataset, however, both models achieved lower accuracy, 14%. On the French dataset, *Meltemi-7B* scored 16%, while *Krikri-8B* scored 17%. On the other hand, the French model *Lucie-7B* showed its best score on the English dataset with 13%, followed by the French dataset with 11%. The model performs poorly on the Greek dataset due to a lack of prior training in that language. Finally, *Mistral Nemo* shows fair performance on the French dataset with an accuracy of 25%, whereas in English it achieved 19%. Despite the fact that it has not been trained for Greek, the model scored 15%; we hypothesize that its multilingual Tekken tokenizer provides better support for Greek than other tokenizers.

4.2 Performance by cardinal and intercardinal directions

In this section, we investigate whether the type of direction – cardinal (e.g., north) versus intercardinal (e.g., northwest) – affects the LLM's responses to benchmark questions. Our results indicate that reasoning involving cardinal directions is more accurate than reasoning involving intercardinal directions; e.g. even in the highest scoring overall, *GPT-5.2*, the accuracies are 84% and 81% respectively[6]. As illustrated in Figure 2 in the Appendix, high accuracy (above 43%) is also observed across all languages for the larger models (*GPT-5.2*, *Kimi-k2-0905*, *Llama-3.3-70B*, *Deepseek V3.2*) with cardinal directions. With intercardinal directions, accuracy is high only for the *GPT-5.2* model as already observed; for the remaining large models, the maximum score achieved is 40% (*Kimi-k2-0905*). This agrees with the findings of [1, 2] for the English dataset. The *Mistral Nemo* model shows the worst results for intercardinal directions; its accuracy across all languages is close to zero (0.01-0.02%). Additionally, *EuroLLM-22B* achieved a 0% score on the French intercardinal directions, while the smaller corresponding model (9B) achieved 16%. Regarding the Greek models, *Meltemi-7B* showed better performance in Greek for intercardinal directions than for cardinal ones (19% vs. 14%). The larger Greek model, *Krikri-8B* showed a score of 16% in intercardinal directions for the French dataset, outperforming its results in Greek (12%) and English (9%). Finally, the French model *Lucie-7B* doesn't perform well with cardinal directions, achieving

[6] A similar result has been seen reported for the non-reasoning model GPT-4.5 in [2].

only 8-10% across all languages. However, it reasons about intercardinal directions more accurately, although it achieves 0% for the Greek dataset, which is expected as the model hasn't been trained on it.

4.3 Performance by question template

The syntactic complexity of a question template is an important factor in the accuracy of an LLM's response. As can be seen in Figure 3 in the Appendix, the easiest template for the LLMs is T5, with the strongest model being *GPT-5.2* which achieved 100% across all languages. The hardest template is T4. Only *GPT-5.2* could achieve high scores in T4 for the French and Greek datasets, with 77% and 75%, respectively. However, for the English dataset, it scored only 8%. Our speculation above that the adaptive reasoning mechanism of *GPT-5.2* may not be working in English's favour may again be explanation for this perhaps surprising difference. The difficulty of T4, compared to other templates, stems from conceptual/semantic confusion introduced by the phrase "road which runs from [one direction] to [another]" since the LLM must distinguish between the agent's movement and the additional information regarding the road's orientation. It is also interesting to note that for T4, *Mistral Nemo* scored 14% in French, 12% in Greek and 15% in English – scoring better than models with a significantly higher number of parameters. Templates T1 and T5 have similar questions, the only difference being the environment: T1 uses "lake", whereas T5 uses "island" and "sea". Additionally, T2 and T6 are more complex extensions of T1 and T5, respectively. Our experiments across all languages show two main results: first, models understand simpler questions (T5, T1) better, and second, models understand the terms "sea shore"/"island" better than "lake shore"/"lake". The reason might be that, for the island question, when you are on the east shore, the sea is to the east, whereas for the lake it is to the west; the latter is arguably the more complicated inference.

4.4 Performance by locomotion and person form

The accuracy across all languages does not change drastically for any particular type of locomotion. Overall, the best performance across all languages was achieved for "hiking" while "riding" performed slightly worse. Similarly, the grammatical person form does not significantly impact LLM performance. The highest accuracy was observed with the second person singular "You" while the third person singular "She"/"He" yielded the lowest results. Nevertheless, the performance difference between them remains negligible. This is in agreement with [1].

5 Future work

We plan to develop benchmark datasets for more European languages (e.g., German and Spanish), and to test our models with very low resource languages such as Basque, as well as right-to-left languages (e.g. Arabic or Hebrew), and also logographic systems such as Chinese. We would also like to experiment with a broader range of reasoning models such as Deepseek-R1 and Phi-4. Given the current poor performance of small models, we would like to try enhancing them using fine-tuning or in context learning. Finally, we would like to consider other spatial reasoning tasks.

References

1 A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions (short paper). In *COSIT 2024*, 2024. doi:10.4230/LIPIcs.COSIT.2024.28.

2 A. G. Cohn and R. E. Blackwell. Evaluating the ability of large language models to reason about cardinal directions, revisited, 2025. Workshop QR@IJCAI 2025. arXiv:2507.12059.

3 A. Petrov et al. Language model tokenizers introduce unfairness between languages, 2023. arXiv:2305.15425.

4 Bryan Li et al. Eliciting better multilingual structured reasoning from LLMs through code, 2024. arXiv:2403.02567.

5 C. Wang et al. Under the shadow of babel: How language shapes reasoning in llms, 2025. arXiv:2506.16151.

6 D. Roussis et al. Krikri: Advancing open large language models for Greek, 2025. arXiv:2505.13772.

7 E. Gardelakos et al. Can large reasoning models reason about spatial relations? In *ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2025.

8 F. Shi et al. Language models are multilingual chain-of-thought reasoners, 2022. arXiv:2210.03057.

9 H. Huang et al. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting, 2023. arXiv:2305.07004.

10 J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models, 2022. arXiv:2201.11903.

11 K. Cobbe et al. Training verifiers to solve math word problems, 2021. arXiv:2110.14168.

12 L. Voukoutis et al. Meltemi: The first open large language model for Greek, 2024. arXiv:2407.20743.

13 Lachlan McPheat et al. Decompsr: A dataset for decomposed analyses of compositional multihop spatial reasoning, 2025. arXiv:2511.02627.

14 Miguel Moura Ramos et al. Eurollm-22b: Technical report, 2026. arXiv:2602.05879.

15 O. Gouvert et al. The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation, 2025. arXiv:2503.12294.

16 Pedro Henrique Martins et al. Eurollm-9b: Technical report, 2025. arXiv:2506.04079.

17 R. E. Blackwell et al. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores, 2024. arXiv:2410.03492.

18 M. Müller and F. Laurent. Cedille: A large autoregressive french language model, 2022. arXiv:2202.03371.

19 S. Skiadopoulos and M. Koubarakis. Composing cardinal direction relations. *Artif. Intell.*, 152(2), 2004. doi:10.1016/S0004-3702(03)00137-1.

A Visualizations of Experimental Results

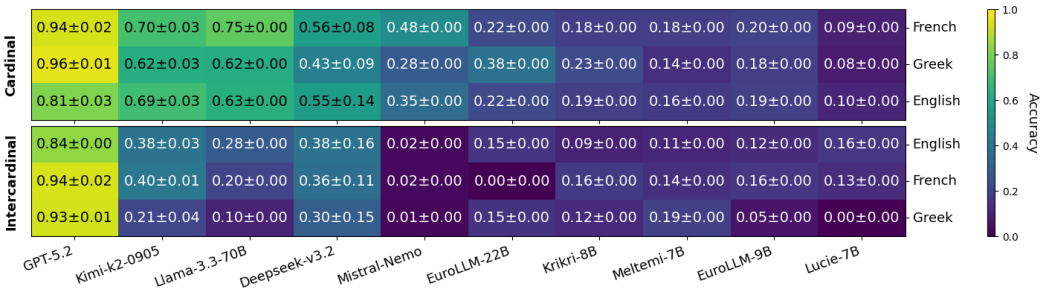


Figure 2 Benchmarking model accuracy: A cross-linguistic comparison of cardinal and intercardinal directions.

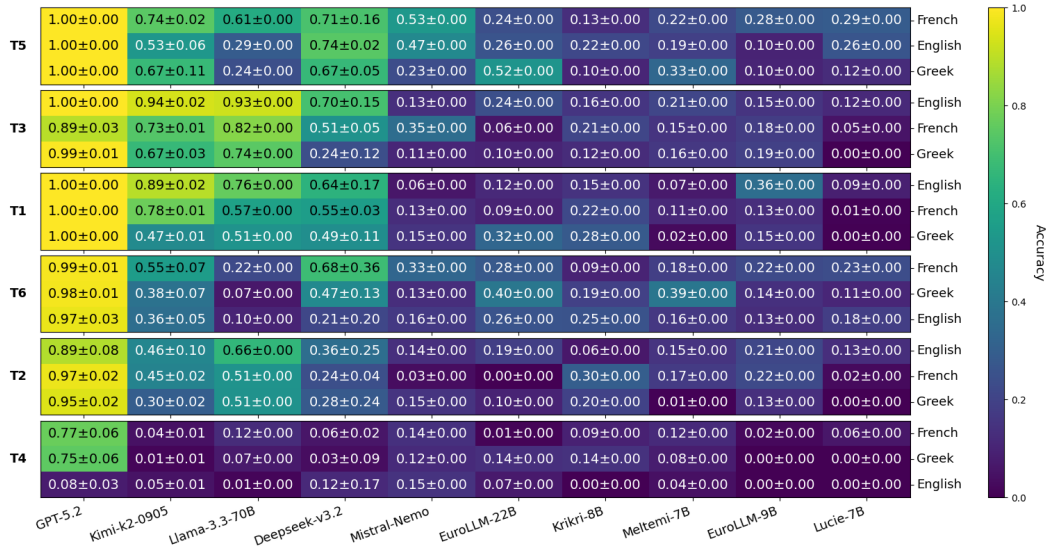


Figure 3 Benchmarking model accuracy: A cross-linguistic comparison by question template.

Structuring the Underwater Environment: A Taxonomy of Aquatic Landmarks

Eva Nuhn[1]
Geoinformatics Group, University of Augsburg, Germany

Sabine Timpf
Geoinformatics Group, University of Augsburg, Germany

Abstract

Effective navigation in aquatic environments relies on the perception and use of environmental cues, or landmarks, that provide spatial orientation and facilitate wayfinding. While much research has focused on terrestrial and urban landmarks, comparatively little is known about underwater landmarks and how divers perceive and use them. Thirty-eight participants completed an online survey and provided information on the use and perception of underwater landmarks. Rock formations and caves, seabed features, and wrecks are the most frequently selected underwater landmarks, with no significant differences observed between recreational and professional divers or among different experience levels. Based on the results of the survey, we develop a data-driven taxonomy of aquatic landmarks comprising three top-level classes: Natural Landmarks, Artificial Landmarks, and Environmental Conditions. The taxonomy provides a structured framework for understanding underwater spatial orientation and serves as a foundation for future research, cognitive mapping studies, and the development of navigational aids in aquatic settings.

2012 ACM Subject Classification Human-centered computing → Empirical studies in HCI; Human-centered computing → Interaction design theory, concepts and paradigms

Keywords and phrases Aquatic navigation, Underwater landmarks, Spatial orientation, Landmark taxonomy, Diver perception

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.19

Category Short Paper

Supplementary Material
Dataset (Research data): `https://doi.org/10.6084/m9.figshare.32604282`

Acknowledgements We would like to thank the student who conducted the survey.

1 Introduction

Effective navigation in aquatic environments relies on the perception and use of environmental cues, or landmarks, that provide spatial orientation and facilitate wayfinding. While extensive research has focused on terrestrial landmarks [15, 26], comparatively little is known about underwater landmarks and how they are perceived by divers. Developing structured conceptual frameworks for underwater landmarks is crucial for applications ranging from recreational diving to marine research, ecological monitoring, and safety operations.

Landmarks in terrestrial and urban environments have been extensively categorised using ontologies and taxonomies, often based on visual salience, spatial location, and functional relevance (e.g., [23, 17, 21, 19]). In aquatic settings, however, landmarks differ fundamentally: the underwater environment presents unique perceptual constraints, including reduced

[1] Corresponding author

© Eva Nuhn and Sabine Timpf;
licensed under Creative Commons License CC-BY 4.0

visibility, variable light conditions, and dynamic water movement. These factors influence which landmarks divers attend to and rely upon for orientation, and they highlight the need for a dedicated, empirically grounded taxonomy of aquatic landmarks.

Frameworks for aquatic environments remain largely unexplored, motivating our effort to fill this gap by developing a data-driven taxonomy of aquatic landmarks. Using the results of an online survey we identify the landmarks most commonly used and map them to structured groups, drawing on a prior framework for landmarks [21] adapting it to the underwater context. This approach incorporates natural landmarks, artificial structures, and environmental conditions, ensuring that the taxonomy reflects both the perceptual realities of diving and the practical needs of navigation. The remainder of the paper is organised as follows: Section 2 reviews existing conceptual frameworks, Section 3 describes the methods employed, Section 4 presents the results, and Section 5 concludes with a discussion and outlook on future work.

2 Conceptual Frameworks for Landmark Representation

Urban landmarks have been categorised in various ways to capture their structural and functional characteristics. Common distinctions include global and local landmarks [23], two-dimensional and three-dimensional landmarks [17], as well as on-route and off-route landmarks, depending on whether they are located at a decision point [14]. Several taxonomies have been proposed to conceptualise the functional roles of landmarks in navigation tasks. Chan and colleagues [5] differentiate between objects serving as visual beacons indicating a nearby en-route target location, objects functioning as associative cues linked to a navigational action, objects acting as visual cues to maintain or regain orientation, and objects used as elements of an environmental reference frame. Similarly, Sadalla and colleagues [20] model the contextual role of landmarks within a hierarchical taxonomy, distinguishing landmarks used in routing tasks, orienting tasks, and memorising tasks. Furthermore, Hansen and colleagues [12] introduce an eight-level taxonomy, with each level describing a specific functional aspect of landmarks within route instructions.

While these classifications provide important structural and functional perspectives, they largely emerge from terrestrial and predominantly urban navigation contexts. As a result, they often assume stable, permanent, and visually salient environmental features [3, 22]. Research on human wayfinding has shown, however, that orientation practices are strongly shaped by environmental conditions, sociocultural needs, and available sensory cues [4, 24, 18]. Research on non-urban landmarks has led to ontology-based approaches tailored to specific environmental contexts. Sarjakoski and colleagues [21] present an empirical study conducted in a national park during both winter and summer seasons. Based on recognised and categorised forest landmarks, they develop a hiking ontology comprising eight landmark groups: structures, passages, trees and parts of trees, waterways, land cover, rocks, signs, and landforms. More recently, Olteanu-Raimond and colleagues [19] propose a lightweight ontology, characterised by relatively low complexity and the absence of formal constraints, designed to support mountain rescue operations by describing different landmark types.

These context-specific approaches demonstrate that wayfinding environments fundamentally influence which features become salient for orientation and how they are conceptualised. This becomes particularly important in marine and underwater environments, where many assumptions underlying terrestrial navigation no longer hold. Unlike urban settings, underwater environments present unique challenges such as low visibility, extreme lighting conditions, dynamic currents, and the unpredictable nature of underwater terrain [11, 10]. Despite

the breadth of existing work, a dedicated conceptual model for water landmarks remains absent from the literature. Tenbrink and Dylla [25] investigate water-related navigation by examining how people talk about space in sailing. Their study shows that sailors rely on specific spatial concepts and specialised jargon such as port and starboard, windward, leeward, or into the wind instead of everyday directional terms. This highlights that common spatial language is insufficient for describing orientation and movement in sailing contexts [25]. However, they do not propose a formal taxonomy or ontology. Their findings nevertheless indicate that marine navigation relies on spatial concepts and orientation strategies that differ fundamentally from those typically assumed in terrestrial landmark research.

3 Methods

This study investigates which landmarks divers use underwater through an online survey. This section provides an overview of the participants who completed the survey (Section 3.1), describes the procedure (Section 3.2), and summaries the statistical analyses employed (Section 3.3).

3.1 Participants

Thirty-eight participants completed an online survey and provided information on the use and perception of underwater landmarks. We recruited participants through diving schools and diving clubs in Germany and France. All participants provided informed consent for participation prior to accessing the questionnaire.

Of the participants, 29 identified as recreational divers and 9 as professional divers. Among recreational divers, years of diving experience ranged from 1 to 50 years (mean = 16.53, SD = 13.23), with the number of completed dives ranging from 31 to 4000 (mean = 728.72, SD = 1002.34). Professional divers reported between 8 and 35 years of experience (mean = 21.11, SD = 9.96) and between 400 and 5000 completed dives (mean = 2473.33, SD = 1893.52).

Six participants are classified as beginners, while the remainder are categorised as advanced or expert divers. Diver classification is based on both years of experience and total logged dives, as some divers may have started years ago but completed relatively few dives. Beginners have 1–5 years of experience or fewer than 100 dives, advanced divers have 6–10 years and at least 100 dives, and experts have more than 10 years and at least 100 dives.

3.2 Procedure

We employ an online survey to investigate which landmarks divers perceive and utilise in aquatic environments. We implemented the questionnaire using SoSci Survey [13] and made available in German and French. A native speaker translated and reviewed the French version. The university's ethics committee approved the procedure. We sent invitation emails containing the survey link in April 2024.

Core questions on use and perception of underwater landmarks asked participants to select from predefined landmark groups including rock formation/caves, distinctive marine fauna, distinctive marine flora, seabed features, corals or sponges, coast, natural coastal landmarks, artificial coastal landmarks, floating buoys, marked dive routes, wrecks, underwater installations, and guide ropes. Participants could also report additional landmarks with an open-ended response.

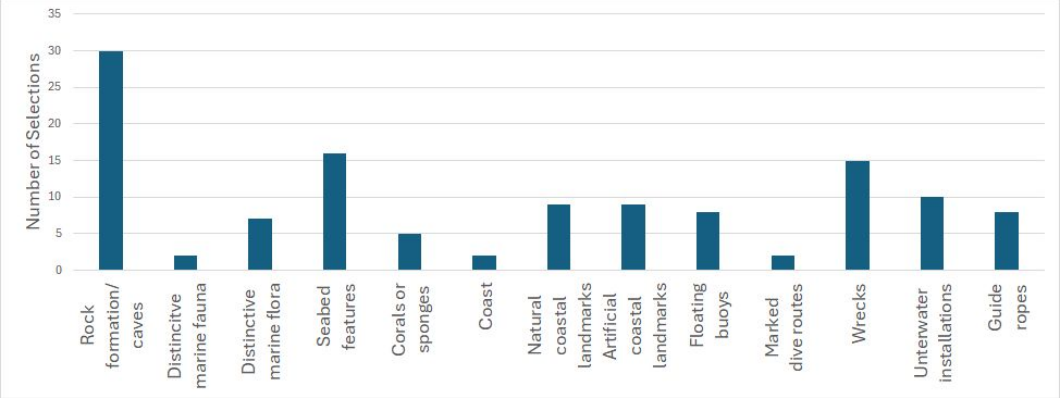


Figure 1 Selections of landmark groups.

3.3 Statistical Analyses

To examine whether the frequency of selection differs across landmark groups within participants, we apply Cochran's Q test [6]. This test is appropriate for detecting overall differences in proportions across more than two related binary variables. When the omnibus test indicates significant differences, post hoc pairwise comparisons are performed using McNemar's tests [16] with Bonferroni adjustment [1] to control for inflated Type I error due to multiple testing.

To explore potential differences between recreational and professional divers in the selection of individual landmarks, we employ Fisher's exact tests [8]. This test is chosen because of small sample sizes and sparse cell counts, which violate the assumptions of chi-square tests. Odds ratios are calculated to estimate the magnitude and direction of group differences. Because multiple landmark groups are tested independently, false discovery rate (FDR) correction [2] is applied to adjust p-values and reduce the risk of Type I error inflation. Finally, to examine associations between diver experience levels and landmark selection across groups with more than two groups, we use the Fisher–Freeman–Halton exact test [9]. This extension of Fisher's exact test is suitable for contingency tables larger than 2×2 and is particularly appropriate for small samples and uneven distributions. All statistical analyses are conducted using an alpha level of 0.05 to determine significance. This threshold is applied consistently across all tests to maintain uniform criteria for evaluating results.

4 Results

In this section we investigate the details by investigating the frequency and patterns of underwater landmark selection (Section 4.1). Furthermore, we propose a data-driven taxonomy of aquatic landmarks (Section 4.2).

4.1 Frequency of Underwater Landmark Selections by Divers

Results indicate that rock formations and caves (n = 30), seabed features (n = 16), and wrecks (n = 15) are the most frequently selected landmarks (Figure 1). Less frequently selected landmarks included distinctive marine fauna (n = 2), the coast (n = 2), and marked dive routes (n = 2). Participants selected an average of 4.10 landmark groups (SD = 1.77).

Table 1 Taxonomy of Underwater Landmarks.

Class	Subclass	Landmark Type
Natural	Geological Landmarks	Cliff; Rock formations; Caves; Holes
	Seabed Cover	Sandy area; Stony area
	Marine Flora	No details given
	Marine Fauna	Fish; Octopus; Moray eels; Corals; Sponges
Artificial	Structures	Wrecks; Submerged car; Underwater installation; Statues; Mine; Wooden cross
	Floating Structures	Boats; Floating buoys
	Navigation Aids	Guide ropes; Spools; Arrows
Environmental	Water Conditions	Depth; Tides; Currents; Visibility
	Celestial	Sun

Cochran's Q test revealed significant differences in selection frequencies across landmark groups ($Q(12) = 107.71$, $p < .001$). Post hoc Bonferroni-adjusted pairwise McNemar tests indicated that rock formations and caves are selected significantly than the other landmarks. Beyond this dominant group, only a few pairwise differences remain significant: distinctive marine fauna versus seabed features, seabed features versus coast, seabed features versus marked dive routes, and marked dive routes versus wrecks.

We observe no significant differences between recreational and professional divers or among different experience levels. Divers did not differ in their selection of underwater landmarks. All associations, tested with Fisher's exact and Fisher–Freeman–Halton exact tests and corrected for multiple comparisons using FDR, are non-significant (all $p > 0.05$).

We analyse open-ended responses thematically. Participants reported using permanent or distinctive landmarks such as buoys, sunken cars, statues, boats, or wooden crosses, especially in areas with low tidal variation or in narrow channels, where proximity to the banks aids navigation. They highlighted visually salient and memorable landmarks, including contrasting sandy and rocky zones, coral formations, and holes or crevices where octopuses and moray eels hide. Environmental conditions, particularly poor visibility in lakes, influenced participants' ability to recognise surroundings, prompting them to follow the group leader closely and rely on navigation aids such as spools and arrows in wrecks or caves. Some also mentioned sunlight and current direction as navigational cues, and certain specific sites, such as the Torpedoversuchsanstalt Tollensee, are referenced as landmarks.

4.2 A Data-Driven Taxonomy of Aquatic Landmarks

Based on the results of the survey, we develop a data-driven taxonomy of aquatic landmarks (Table 1). Analyses reveal no perceptual differences between recreational and professional divers, nor among divers with varying levels of experience. Consequently, we develop a single unified taxonomy for all participants.

The final taxonomy for aquatic landmarks consists of three top-level classes: Natural Landmarks, Artificial Landmarks, and Environmental Conditions, subdivided into thematic subclasses (e.g., Geological Landmarks, Marine Flora). Each subclass contains specific landmark types (e.g., Cliff, Rock formation, Caves) (Table 1). To ensure a data-driven taxonomy, we include only terms that participants were exposed to or explicitly named.

The following paragraph details how each subclass is defined and adapted based on survey responses and environmental considerations. We adopt the groups of Sarjakoski and colleagues [21] (structures, passages, trees, water, land cover, rocks, signs, and landforms)

and map the observed landmarks to these groups wherever appropriate. Landmarks that are rarely selected by participants (distinctive marine fauna, n = 2; coast, n = 2; marked dive routes, n = 2) are considered for inclusion with caution, and coast and marked dive routes are excluded due to low selection frequency. However, open-ended responses frequently referenced fish, octopus, moray eels, and corals, motivating the addition of a *Marine Fauna* subclass. Because terrestrial land cover does not translate directly to underwater environments, we refine this group into a *Seabed Cover* subclass. To capture non-living solid formations, we introduce a *Geological Landmarks* subclass encompassing both landforms and rocks. Recognising the absence of trees in marine environments, we incorporate a *Marine Flora* subclass to include landmarks such as seagrass, kelp, and algae fields. Although participants selected the class Natural Coastal Landmarks, they did not provide any specific details, and therefore no corresponding landmark type is included in the taxonomy. Within artificial landmarks, we divide the original Structures class into *Structures* (fixed man-made objects, e.g., wrecks, statues, submerged vehicles) and *Floating Structures* (e.g., boats, floating buoys) to account for differing dynamics and mapping properties. Since participants did not report conventional signs such as informational boards or warning markers, we do not include a Signs class. All navigational cues, including markers such as spools and arrows, are included in a newly defined *Navigation Aids* subclass. Finally, we create two additional subclasses to capture environmental and dynamic conditions: *Water Conditions*, which includes depth, current, and visibility, and *Celestial Conditions*, capturing landmarks such as the sun.

5 Discussion and Outlook

The taxonomy provides a structured framework for understanding underwater spatial orientation and serves as a foundation for future research. Rock formations and caves, seabed features, and wrecks emerged as the most salient landmarks. Objects such as the coast, marked dive routes, and distinctive marine fauna are rarely selected. Interestingly, we find no statistically significant differences between recreational and professional divers, nor among divers with varying experience levels. This suggests that the identification of key underwater landmarks may be largely universal, governed by perceptual and environmental constraints rather than training or skill. This finding also justifies the development of a single, unified taxonomy for all divers, rather than separate frameworks for different user groups.

Some of the adapted groups from Sarjakoski et al. [21] required refinement to suit the underwater context: terrestrial land cover became Seabed Cover, trees are replaced with Marine Flora, and Geological Landmarks are created to encompass rocks and non-living solid formations. Artificial Landmarks are divided into Structures and Floating Structures, and Navigation Aids are added to capture orientation cues. Environmental conditions, including Water Conditions and Celestial Conditions, are also included to reflect dynamic factors affecting navigation. These adaptations illustrate the importance of contextualising landmark groups to the specific affordances of the underwater environment.

While the taxonomy is data-driven and grounded in participant observations, some groups are informed by open-ended responses rather than frequency counts, which may introduce subjective interpretation. Additionally, the sample size is modest, particularly for professional divers, which may have limited the ability to detect subtle experience-based differences. Future studies could expand the participant pool, examine landmarks across different aquatic environments, or investigate how dynamic factors such as currents, visibility, or moving fauna influence navigational strategies.

Future work will explore underwater landmark modalities beyond vision, with a particular focus on acoustics. Sound is the most efficiently transmitted radiation underwater, creating a complex "soundscape" of natural, biological, and anthropogenic sources that may serve as navigational cues for marine life [7]. Investigating how divers could perceive or utilise these acoustic landmarks may provide new insights into orientation strategies and inform the design of non-visual navigational aids.

Overall, this work contributes both empirical and conceptual insights. It confirms that underwater navigation relies on landmarks, highlights the consistency of landmark perception across diver experience, and provides a structured, data-driven taxonomy that can serve as a foundation for future research, cognitive mapping studies, and the development of navigational aids in aquatic settings.

References

1 Hervé Abdi. Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3(01):1–9, 2007.

2 Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, December 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.

3 Gary Burnett, Darren Smith, and Andrew May. Supporting the navigation task: Characteristics of 'good' landmarks. In Margaret Hanson, editor, *Contemporary Ergonomics 2001*, pages 506–512. CRC Press, 0 edition, 2001. doi:10.1201/b12798-69.

4 David Caduff and Sabine Timpf. On the assessment of landmark salience for human navigation. *Cognitive Processing*, 9(4):249–267, 2008. doi:10.1007/s10339-007-0199-2.

5 Edgar Chan, Oliver Baumann, Mark A Bellgrove, and Jason B Mattingley. From objects to landmarks: the function of visual location information in spatial navigation. *Frontiers in psychology*, 3:304, 2012. doi:10.3389/fpsyg.2012.00304.

6 William G Cochran. The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266, 1950. doi:10.2307/2332378.

7 A John R Cotter. The "soundscape" of the sea, underwater navigation, and why we should be listening more. *Advances in fisheries science*, 50:451–471, 2008. doi:10.1002/9781444302653.ch19.

8 Ronald Aylmer Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1934.

9 GH Freeman and John H Halton. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38(1/2):141–149, 1951. doi:10.2307/2332323.

10 Tianyi Fu, Bingxiong Wang, Yang Liu, Runtong Ai, Jincun Liu, Cong Wang, and Daoliang Li. Design, perception, planning and control for underwater cleaning robotics: Progress, challenges and trends. *Journal of Ocean Engineering and Science*, 11(2):614–637, 2026. doi:10.1016/j.joes.2025.12.019.

11 M. Grimaldi, D. Nakath, M. She, and K. Köser. Investigation of the challenges of underwater-visual-monocular-slam. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:1113–1121, 2023. doi:10.5194/isprs-annals-x-1-w1-2023-1113-2023.

12 Stefan Hansen, Kai-Florian Richter, and Alexander Klippel. Landmarks in OpenLS—a data structure for cognitive ergonomic route directions. In *International Conference on Geographic Information Science*, pages 128–144. Springer, 2006. doi:10.1007/11863939_9.

13 D. J. Leiner. SoSci Survey (Version 3.5.02) [Computer software]. https://www.soscisurvey.de, 2026. Last Accessed February 23, 2026.

14 Kristin L. Lovelace, Mary Hegarty, and Daniel R. Montello. Elements of good route directions in familiar and unfamiliar environments. In Christian Freksa and David M. Mark, editors, *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. COSIT 1999*, volume 1661 of *LNCS*, pages 65–82. Springer, 1999. doi:10.1007/3-540-48384-5_5.

15 Kevin Lynch. *The Image of the City*. Publication of the Joint Center for Urban Studies. M.I.T. Press, 33rd print ed. edition, 1960.

16 Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947. doi:10.1007/BF02295996.

17 Pierre-Emmanuel Michon and Michel Denis. When and Why Are Visual Landmarks Used in Giving Directions? In Daniel R. Montello, editor, *Spatial Information Theory: Foundations of Geographic Information Science: International Conference, COSIT 2001, Morro Bay, CA, USA, September 19-23, 2001: Proceedings*, volume 2205 of *Lecture Notes in Computer Science*, page 502. Springer, 2001. doi:10.1007/3-540-45424-1_20.

18 Eva Nuhn, Kai Hamburger, and Sabine Timpf. Mapping olfactory cues for wayfinding – a theoretical approach and an empirical study. *Journal of Location Based Services*, 19(1):78–102, 2024. doi:10.1080/17489725.2024.2371298.

19 Ana-Maria Olteanu-Raimond, Mattia Bunel, Catherine Dominguès, Cécile Duchêne, Laurence Jolivet, Sébastien Mustière, Marie-Dominique Van Damme, Olivier Favre, and Véronique Gendner. A lightweight ontology for landmarks to assist rescue in mountainous areas. *Advances in Cartography and GIScience of the ICA*, 4:15, 2023. doi:10.5194/ica-adv-4-15-2023.

20 Edward K Sadalla, W Jeffrey Burroughs, and Lorin J Staplin. Reference points in spatial cognition. *Journal of experimental psychology: human learning and memory*, 6(5):516, 1980. doi:10.1037/0278-7393.6.5.516.

21 Tiina Sarjakoski, Pyry Kettunen, Hanna-Marika Halkosaari, Mari Laakso, Mikko Rönneberg, Hanna Stigmar, and Tapani Sarjakoski. Landmarks and a hiking ontology to support wayfinding in a national park during different seasons. In Martin Raubal, David M. Mark, and Andrew U. Frank, editors, *Cognitive and Linguistic Aspects of Geographic Space: New Perspectives on Geographic Information Research*, pages 99–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-34359-9_6.

22 Allison L Thomas Scrivner-Limbaugh. *Real-scene contextual cueing: why trees are better than cats*. PhD thesis, University of Alabama, Department of Psychology, 2015.

23 Sibylle Steck and Hanspeter Mallot. The Role of Global and Local Landmarks in Virtual Environment Navigation. *Presence Teleoperator. Virtual Env.*, 9, 1998. doi:10.1162/105474600566628.

24 Paul Symonds, David HK Brown, and Valeria Lo Iacono. Exploring an absent presence: Wayfinding as an embodied sociocultural experience. *Sociological Research Online*, 22(1):48–67, 2017. doi:10.5153/sro.4185.

25 Thora Tenbrink and Frank Dylla. Sailing: Cognition, action, communication. *Journal of Spatial Information Science*, 15(2017):3–33, 2017. doi:10.5311/JOSIS.2017.15.350.

26 Demet Yesiltepe, Ruth Conroy Dalton, and Ayse Ozbil Torun. Landmarks in wayfinding: A review of the existing literature. *Cognitive Processing*, 22(3):369–410, 2021. doi:10.1007/s10339-021-01012-x.

GeoKG for GeoFM: Neuro-Symbolic Integration of Semantically-Enriched Geospatial-Temporal Data

Kai Moltzen
AIX, Leuphana University Lüneburg, Germany

Debayan Banerjee
AIX, Leuphana University Lüneburg, Germany
HITeC e.V., Hamburg, Germany

Julian Burmester
AIX, Leuphana University Lüneburg, Germany

Anna Ehrenberg
AIX, Leuphana University Lüneburg, Germany

Martin Kohler
AIX, Leuphana University Lüneburg, Germany

Cedric Möller
University of Hamburg, Germany

Jann Pfeifer
AIX, Leuphana University Lüneburg, Germany

Tilahun Abedissa Taffa
AIX, Leuphana University Lüneburg, Germany

Hanna Marlene Schulz
AIX, Leuphana University Lüneburg, Germany

Aida Usmanova
AIX, Leuphana University Lüneburg, Germany

Patrick Westphal
AIX, Leuphana University Lüneburg, Germany
HITeC e.V., Hamburg, Germany

Ricardo Usbeck
AIX, Leuphana University Lüneburg, Germany
HITeC e.V., Hamburg, Germany

Abstract

Solving complex real-world problems requires integrating heterogeneous data. Although recent Geospatial Foundation Models (GeoFMs) excel at pixel-level analysis, they make only limited use of the geographic information contained in natural language texts and therefore often lack the semantic context needed for complex geospatial reasoning. This vision paper argues for reconceiving unstructured texts as a core component of a neuro-symbolic framework that populates a geospatial-temporal knowledge graph (GeoKG). We outline an architecture in which modality-specific extractors produce multi-granular footprints that are fused into a GeoKG, synchronizing symbolic representations with embeddings, geometries, and grid cells. This GeoKG supports geospatial reasoning and provides enriched training data for truly multimodal GeoFMs, for which we pose a set of research challenges.

2012 ACM Subject Classification Information systems → Ontologies; Computing methodologies → Spatial and physical reasoning; Computing methodologies → Information extraction; Computing methodologies → Learning latent representations

Keywords and phrases GeoKG, Foundation Models, GeoFM, neuro-symbolic Data Integration

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.20

Category Short Paper

Funding The authors acknowledge financial support from the Federal Ministry for Economic Affairs and Energy (BMWe) of Germany for the project CoyPu (project number 01MK21007G), and from the Ministry of Research and Education (BMBF) for the project "RESCUE-MATE: Dynamische Lageerstellung und Unterstützung für Rettungskräfte in komplexen Krisensituationen mittels Datenfusion und intelligenten Drohnenschwärmen" (FKZ 13N16844). This work is also supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259, and further funded by the German Federal Ministry for Research, Technology, and Space (BMFTR) under Grant No. 13N17698 (PROVIDER). This publication was produced as part of the project "Leuphana AI Campus (LAICA)", funded by the Foundation for Innovation in Higher Education.

Declaration on GenAI/LLM use During the preparation of this work, the authors used generative models hosted by the ChatAI service of the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) and Grammarly to check grammar and spelling, paraphrase and reword, and improve writing style. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

© Kai Moltzen, Debayan Banerjee, Julian Burmester, Anna Ehrenberg, Martin Kohler, Cedric Möller, Jann Pfeifer, Tilahun Abedissa Taffa, Hanna Marlene Schulz, Aida Usmanova, Patrick Westphal, and Ricardo Usbeck;
licensed under Creative Commons License CC-BY 4.0

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 20; pp. 20:1–20:9

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

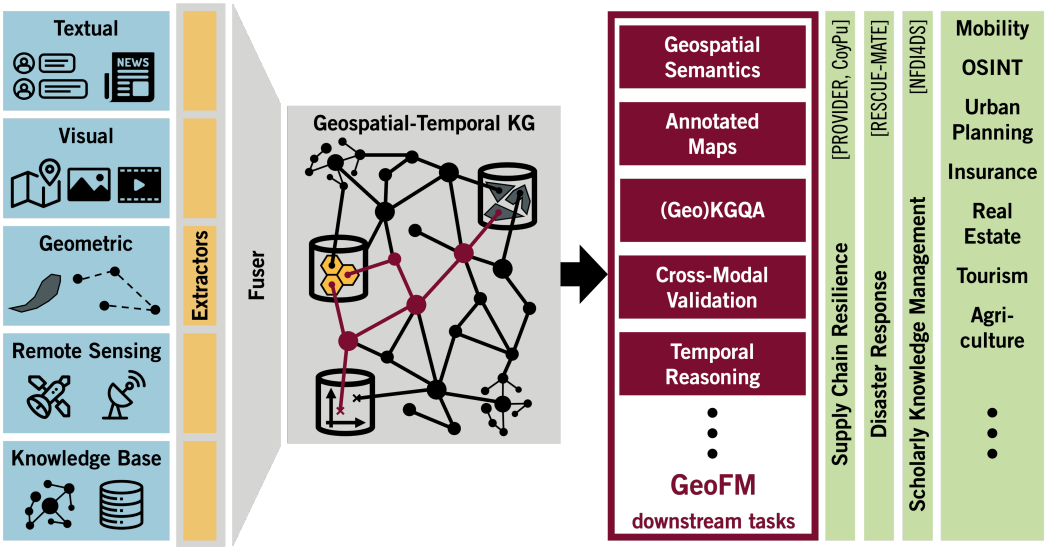


Figure 1 Vision architecture **GeoKG for GeoFM**. We integrate geospatial-temporal information extracted from textual, visual, RS, and geometric data sources via extractor modules and fuse the results into an ontology-based multimodal GeoKG, synchronized with cell, geometric, raster, and vector stores. The red links illustrate cross-modal alignment by geospatial relation, connecting grid cells, geometries, and embedding vectors within the GeoKG. The resulting enriched, spatially and temporally aligned data backbone is intended as training data for GeoFMs and diverse applications.

Addressing complex real-world challenges across domains, from disaster response and supply chain resilience to scholarly knowledge management, requires the integration of heterogeneous data. As we were able to demonstrate in interdisciplinary research consortia such as CoyPu, RESCUE-MATE, PROVIDER, and NFDI4DS[1], decision support in these domains relies on aligning diverse modalities: unstructured text, remote sensing (RS) imagery, structured geometrical data, and entire knowledge bases. However, a fundamental barrier remains: while these modalities can describe the same physical *reality*, they differ significantly in spatial granularity, temporal resolution, and semantic structure.

To bridge this gap, raw data must be geographically, temporally, and semantically aligned and enriched. Throughout, we use the term alignment in this data-centric sense: co-locating heterogeneous representations of the same place or event in a shared spatial, temporal, and semantic reference, and distinguish it from the separate notion of aligning model behavior with human values, which we discuss in Section 3.2. Semantic data enrichment is an established process for enhancing FAIRness [36] and providing meaning and context for data by means of an ontology [9, 35]. In this work, we envision a novel framework that extends enrichment beyond entity and relation extraction. Source- and modality-specific extractor modules derive the geospatial-temporal dimension of the input data at different granularities. By fusing enriched information into a multimodal geospatial-temporal Knowledge Graph (GeoKG), various input data can be semantically aggregated, filtered, and further processed using reliable spatiotemporal footprints.

[1] `https://coypu.org/en/`; `https://www.rescue-mate.de/`; `https://projekt-provider.de/`; `https://www.nfdi4datascience.de/`

Throughout this paper, we return to one running example to illustrate this framework: a flood that disrupts a supply chain. News reports describe inundated areas and explain how rescue services and organizations react to it based on past events; RS imagery delineates the flooded extent; in-situ sensors report rising water levels at regular timestamps; and geometric data locate the affected roads, warehouses, and supplier sites. An operational question such as "which product deliveries and transport routes are at risk, and why?" can only be answered by bringing all of these sources into a common spatial, temporal, and semantic frame.

In this vision paper, we argue that the current trajectory of Geospatial Artificial Intelligence (GeoAI) [9] is constrained by the limited role that geographic information from texts plays in the emerging generation of Geospatial Foundation Models (GeoFMs). While recent surveys indicate a surge in multimodal GeoFMs [37, 18], these models are predominantly trained on RS data. When text is included, it is often limited to vision-language tasks like image captioning or visual grounding [37]. Consequently, current systems struggle to perform complex semantic geospatial reasoning tasks such as question answering (GeoQA) [26, 27] that require understanding the "why" and "how" of an event – for instance, for whom a particular road is impassable and how that propagates to downstream deliveries in our flood scenario. Such information is rarely visible in pixels alone but abundant in textual narratives. Natural language (NL) texts can therefore enhance GeoFMs in reasoning, in the retrieval of geographically similar documents, and as a valuable source of information about events and place semantics [23].

To enable genuinely multimodal GeoFMs that reason across the full spectrum of spatial, temporal, and textual evidence, this vision paper argues for reconceiving texts containing geographical information as an integral component of geospatial-temporal, multimodal Knowledge Graphs (KGs), as shown in Figure 1. Such GeoKGs can in turn serve as a rich source of training data for the next generation of GeoFMs. We present concrete research questions for the community to discuss, and look forward to receiving feedback, as our framework will serve as the foundation for future research.
Our contributions are as follows:

1. **Unified geospatial-temporal extraction and fusion as a fundamental step:** Our extractors first derive information about the geospatial extent (e.g., polygons, Discrete Global Grid System (DGGS) cells) and temporal (meta-)data (e.g., validity intervals, event timestamps) of the diverse inputs, then fuse and map it to a central GeoKG.
2. **Neuro-symbolic data integration of all core concepts:** By coupling structured, symbolic KG construction with dense-vector embeddings and raster maps, our framework integrates all core concepts of spatial information [13] and supports semantic similarity search [28, 8] across modalities while preserving the logical rigor of ontological reasoning.
3. **Data backbone for GeoFMs:** The GeoKG provides a structured training ground of richly annotated, geospatially and temporally aligned textual descriptions linked to multi-granular spatial footprints. We expect this grounding to enhance truly multimodal GeoFMs and improve tasks such as GeoQA, reasoning, and cross-modal retrieval.

2 Neuro-Symbolic Geospatial-Temporal Architecture

The architecture in Figure 1 ranges from diverse input sources to domain-specific use cases. At its core, the GeoKG is designed for geospatial multimodality, interlinking KG nodes with textual descriptions to standardized grid cells, geometric shapes, or dense vector embeddings.

2.1 Input Sources

Geospatial-temporal data can be captured in various forms [17]. NL textual descriptions take the form of news articles, social media posts, or more formalized modes such as encyclopedic descriptions; in our running example, these include the news reports describing the flood and its consequences for the supply chain. These sources also provide visual modalities, such as images, videos, and illustrated maps. Street-view and drone imagery, as well as RS, are essential sources within the geospatial domain, with the latter typically further categorized into, e.g., optical RGB, spectral, or SAR [37]. Geometric data types like points, polylines, and (multi-)polygons represent, e.g., geographic regions, trajectories, or building footprints, such as the routes and warehouse outlines in the flood scenario.

Text is a prominent modality in KGs in general, but geographic and place information in texts is rarely captured in GeoKGs beyond entity linking. We suppose this is because such spatial language must first be expensively extracted and interpreted. Consequently, we rely on an explicit ingestion layer as a fundamental step for our GeoKG.

2.2 Ingestion Layer

Modality-specific extractors are not only necessary for aligning the various data types, but also for enriching and interpreting the information for representation in the GeoKG. For example, NL descriptions of real-world spatiotemporal phenomena are inherently vague. Therefore, complex location descriptions must be interpreted reliably, and the extents of, e.g., the impact area of natural disasters, estimated based on the context [19, 14]. In the flood scenario, an extractor must turn a phrase such as "the lower river valley is under water" into a concrete, multi-granular footprint. Other examples include the spatial and temporal aggregation of social media posts or narrative extraction [6]. After extraction, all information is aligned and enriched semantically within the fuser module, thereby mapping it to the GeoKG ontology.

2.3 Data Layer: GeoKG

We populate a GeoKG by reusing existing ontologies [4] (e.g., KWG-Ont [29], PROV-O [1], or CoyPu [31]). We expect to further develop novel ontologies to represent heterogeneous spatial data types [13, 27] with different degrees of granularity, temporality, and vagueness. Among those data types, existing GeoKGs focus on objects, events, and networks, but rarely consider fields. To enable fully four-dimensional, spatiotemporal 'hyperobjects' [5], however, fields indicating presence in space-time constitute a promising means. Fields enable us to consider the temporal dimension of events more adequately than with mere object-timestamps, as common in spatiotemporal KGs [22]. Besides geometric shapes like points, lines, or polygons stored in well-known text (WKT) strings, we will thus implement synchronization with standardized cells within DGGSs [21], and dense vector embeddings for downstream usage.

A DGGS partitions the Earth into a hierarchy of nested, approximately equal-area cells, each with a stable identifier [25, 21], and this plays a central role in our design for three reasons. First, a cell identifier acts as a modality-independent join key: a flooded pixel in an RS scene, a polygon describing an inundated district, and a sentence georelated to a disaster area can all be attached to the same cell and thereby related without committing to a single geometry. Second, the cell hierarchy provides multi-granular footprints, so that coarse, vague textual descriptions and fine-grained sensor readings can coexist and be reconciled at a shared resolution. Third, aggregation over cells – for example, collecting all NL labels and embeddings that refer to a given area within a time window – yields compact, queryable

summaries that function as supervision for downstream models. In our running example, the affected region is thus represented as a set of cells that links the news- and RS-derived event, extent, and the geometric footprints of the threatened warehouses and routes, while the corresponding text embeddings remain attached for similarity search. Fundamentally, this synchronization allows us to align arbitrary spatiotemporal representations, extending the foundational work on combining KGs and DGGS in the KnowWhereGraph project [40]. Therefore, our GeoKG constitutes a valuable source of training data for truly multimodal GeoFMs (see Section 3.1). Technically, this capability is based on the synchronization between a triple store for the KG, geospatial relational databases (DB), and DB extensions for vector embeddings and DGGS cells.

2.4 Interface and Service Layer

The interface layer provides a stable, unified interface to the GeoKG and external services. It offers direct graph access (e.g., SPARQL over named graphs), modality-specific endpoints for spatial queries and data access (e.g., GeoSPARQL), vector similarity search, and access to external resources and interfaces (e.g., routing, weather APIs). On top of the interface layer, the service layer provides reusable capabilities that can be shared across downstream domain applications. Services orchestrate multi-step operations that cannot be expressed as single queries or isolated store operations. Examples include services for geospatial semantics, temporal reasoning, cross-modal validation, geographic visualization, and semantic interfaces like GeoKG Question Answering (GeoKGQA), which incorporates spatial inference and enables discovery and reasoning about geospatial relationships between entities – for instance, answering which suppliers fall within the flooded cells and which routes connect them.

2.5 Domains

Each application implements domain-specific functions while reusing common services for geospatial-temporal grounding. The research project PROVIDER develops a decision-support system for strengthening supply chain resilience, the setting of our running example. Real-time news streams must be processed to extract events and explicitly link them to spatiotemporal footprints, such as facilities, routes, and assets, integrating existing knowledge on dependencies and constraints. This domain, therefore, relies on aligning heterogeneous sources and provenance information, supported by the GeoKG.

The research project RESCUE-MATE provides an exemplary disaster response application, the perspective from which the same flood is monitored as it unfolds. It maintains a digital twin updated from real-time sources (rescue radio traffic, environmental sensors), complemented by social media data to serve as the foundation for decision-making in crises. The shared geospatial-temporal services support aligning incoming observations with the GeoKG and enable aggregating local reports into an operational picture.

NFDI4DS constitutes a research consortium for scholarly knowledge management, for which our framework targets the extraction and linking of geographic aspects of research artifacts and their alignment with a DGGS. NFDI4DS will provide interfaces for geospatial-temporal searching and filtering of scholarly data, e.g., all papers on a certain regularly-flooded region across various research fields in a given time span.

Our framework will support further domains that depend on geospatial-temporal grounding and multimodal data integration, such as urban planning, insurance, real estate, tourism, agriculture, and mobility, all of which require integrating data with authoritative geospatial-temporal reference data.

3 Overarching Perspectives

3.1 Towards (Truly) Multimodal GeoFMs

Recent years have seen a surge in GeoFM publications and performance [18, 37, 16]. Advancements are driven by models such as Prithvi-EO-x [32] or AlphaEarthFoundations [2], which excel at pixel-level tasks, including segmentation and scene classification. However, the former model only considers RS modalities and no text at all; the latter relies mostly on Wikipedia articles with document-level point coordinates. Vision-language GeoFMs largely come down to RS visual grounding and image QA tasks [39, 15]. We observe that geographic information from texts so far plays only a minor role in the current generation of GeoFMs, despite transformer models for NL processing having sparked the recent surge of interest in foundation models in the first place.

We expect that the wealth of unstructured geographic insights captured in NL texts will contribute greatly to the next generation of truly multimodal GeoFMs. Individuals and societies attribute meaning to places, largely expressed through natural language and captured in texts. As Purves and colleagues argue [23], places shall thus be jointly represented by textual attributes and locations, for which the GeoKG depicts an ideal technology to store, align, and semantically enrich place information. Using the NL semantic labels aggregated by DGGS cells within the GeoKG will serve as a much-needed standardization. It constitutes an important step toward adopting GeoFMs, or earth embeddings, to NL processing tasks such as geographic information retrieval [11, 8], for which GeoFMs can then serve as dense prior information.

Returning to our running example, a query for documents about places similar to the flooded district can be answered by combining the textual descriptions and the earth embeddings aggregated per cell. Consequently, mitigation and disaster relief policies from these places might be transferred to the current impact area.

Our proposed architecture makes two contributions: (1) we include extractors to interpret and align geospatial-temporal information before ingestion into a GeoKG, and (2) we design the GeoKG to represent and synchronize different data types, accounting for the heterogeneity of spatial data as an important data source for truly multimodal GeoFMs capable of complex geospatial reasoning and retrieval.

3.2 Representation of Heterogeneous Geographic Data

Geographical data is the translation of the complexity of spatial *realities* into computable representations. Because of this dependence on users' cultures [33], pluralistic values, or legal regulations [7], we acknowledge that committing our framework to responsible AI practices alone is insufficient to address systemic power imbalances [24, 3]. Drawing on feminist AI literature [34], we incorporate principles such as critical assessment of power structures, consideration of environmental impacts, context, and consent [12]. Yoo [38] translates GeoAI-specific principles into a practical framework consisting of guidelines and checklists. The framework draws on contributions from industry, academia, and global governance institutions. It addresses risks that are particular to GeoAI, such as linkage attacks, surveillance, and geographic representational imbalances, resulting in twelve ethical GeoAI axes. In our future work, we aim to test this framework in practice by applying the proposed checklists.

Further, the synchronization and representation across spatial data formats help us to draw on geographical fuzziness as an important source of knowledge rather than as mere noise to be removed. Vague extents such as the referent of a place name or the cognitive

boundary of a region carry meaning of their own [20], and our shared representation preserves this graded structure instead of forcing exact boundaries. Indigenous perspectives on tribal land, e.g., create not only different data, but can also show alternative concepts of spatial representations. The dynamic nature of land-shaping forces, the fractal nature of boundaries, or the uneven distribution of data collection result in inherent uncertainty. Place names are rarely neutral but bound to hierarchies and power through their implicit political nature. Distinguishing, e.g., between coarse cells or clearly defined boundaries, in combination with the semantical richness of textual descriptions in the GeoKG, enables the representation of ambiguous [10] or contested regions [30] to account for this heterogeneity.

4 Conclusions Towards Future Research Agendas

Realizing the proposed framework leads to a set of foundational research questions. On the representation side, we need principled models of time and change, identity across evolving geometries, and a clear semantics of events with potentially fragmented or recurrent time spans. On the learning side, we require embedding spaces that integrate heterogeneous geospatial-temporal modalities, including GeoKG embeddings, and causal or neuro-symbolic representations that capture more than surface-level similarity. On the integration side, a key challenge is bidirectional translation between symbolic and neural representations, namely, when and how to map learned representations back into structured representations and formats used in GIS workflows. Finally, for GeoFMs, we lack consensus on training tasks, supervision signals, and evaluation protocols that reward complex semantic geospatial reasoning, provenance, and auditability, while meeting FAIRness requirements towards heterogeneity, representativeness, privacy, and accountable use in high-stakes settings.

References

1 Khalid Belhajjame, James Cheneyand David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV ontology. W3C recommendation, World Wide Web Consortium (W3C), April 2013. URL: https://www.w3.org/TR/2013/REC-prov-o-20130430/.

2 Christopher F. Brown, Michal Kazmierski, Valerie Pasquarella, and colleagues. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *CoRR*, abs/2507.22291, 2025. doi:10.48550/arXiv.2507.22291.

3 Eleanor Drage, Kerry McInerney, and Jude Browne. Engineers on responsibility: feminist approaches to who's responsible for ethical AI. *Ethics and Information Technology*, 26(1):4, 2024. doi:10.1007/s10676-023-09739-1.

4 Mariano Fernández-López, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. Methodological guidelines for reusing general ontologies. *Data Knowl. Eng.*, 86:242–275, 2013. doi:10.1016/J.DATAK.2013.03.006.

5 Antony Galton. Fields and objects in space, time, and space-time. *Spatial Cogn. Comput.*, 4(1):39–68, 2004. doi:10.1207/s15427633scc0401_4.

6 Junbo Huang and Ricardo Usbeck. Narration as functions: from events to narratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 1–7, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.wnu-1.1.

7 Krzysztof Janowicz, Zilong Liu, Gengchen Mai, Zhangyu Wang, Ivan Majic, Alexandra Fortacz, Grant McKenzie, and Song Gao. Whose Truth? Pluralistic Geo-Alignment for (Agentic) AI. In *SIGSPATIAL 2025*, pages 799–803. ACM, 2025. doi:10.1145/3748636.3760465.

8 Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2(1):29–57, May 2011. doi:10.5311/JOSIS.2011.2.3.

9 Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. Geospatial semantics and linked spatiotemporal data - past, present, and future. *Semantic Web*, 3(4):321–332, 2012. doi:10.3233/SW-2012-0077.

10 Zhe Jiang, Yan Li, Shashi Shekhar, Lian P. Rampi, and Joseph F. Knight. Spatial ensemble learning for heterogeneous geographic data with class ambiguity: A summary of results. In *SIGSPATIAL 2017*, pages 23:1–23:10. ACM, 2017. doi:10.1145/3139958.3140044.

11 Christopher B. Jones and Ross S. Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008. doi:10.1080/13658810701626343.

12 Lauren F. Klein and Catherine D'Ignazio. Data feminism for AI. In *FAccT 2024*, pages 100–112. ACM, 2024. doi:10.1145/3630106.3658543.

13 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *Int. J. Geogr. Inf. Sci.*, 26(12):2267–2276, 2012. doi:10.1080/13658816.2012.722637.

14 Egoitz Laparra and Steven Bethard. A dataset and evaluation framework for complex geographical description parsing. In *COLING 2020*, pages 936–948. ICCL, 2020. doi:10.18653/V1/2020.COLING-MAIN.81.

15 Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote. Sens.*, 58(12):8555–8566, 2020. doi:10.1109/TGRS.2020.2988782.

16 Gengchen Mai, Weiming Huang, Jin Sun, and colleagues. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Trans. Spatial Algorithms Syst.*, 10(2):11, 2024. doi:10.1145/3653070.

17 Gengchen Mai, Xiaobai Angela Yao, Yiqun Xie, Jinmeng Rao, Hao Li, Qing Zhu, Ziyuan Li, and Ni Lao. SRL: towards a general-purpose framework for spatial representation learning. In *SIGSPATIAL 2024*, pages 465–468. ACM, 2024. doi:10.1145/3678717.3691246.

18 Valerio Marsocci, Yuru Jia, Georges Le Bellier, and colleagues. Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–43, 2025. doi:10.1109/MGRS.2025.3628194.

19 Kai Moltzen, Junbo Huang, and Ricardo Usbeck. LLM Agents for Georelating - A New Task for Locating Events. In *SIGSPATIAL 2025*, pages 277–280. ACM, 2025. doi:10.1145/3748636.3762733.

20 Daniel R. Montello, Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl. Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, 3(2-3):185–204, September 2003. doi:10.1080/13875868.2003.9683761.

21 OGC. Discrete global grid systems abstract specification, 01.08.2017. URL: https://docs.ogc.org/as/15-104r5/15-104r5.html.

22 Philipp Plamper, Hanna Köpcke, and Anika Groß. A survey on spatio-temporal knowledge graph models. *CoRR*, abs/2512.16487, 2025. doi:10.48550/arXiv.2512.16487.

23 Ross S. Purves, Stephan Winter, and Werner Kuhn. Places in information science. *J. Assoc. Inf. Sci. Technol.*, 70(11):1173–1182, 2019. doi:10.1002/asi.24194.

24 Paola Ricaurte and Mariel Zasso. AI, ethics, and coloniality: A feminist critique. In *What AI Can Do: Strengths and Limitations of Artificial Intelligence*, pages 39–58. Chapman and Hall/CRC, New York, NY, 2023. doi:10.1201/b23345.

25 Kevin Sahr, Denis White, and A. Jon Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003. doi:10.1559/152304003100011090.

26 S. Scheider, A. Ballatore, and R. Lemmens. Finding and sharing GIS methods based on the questions they answer. *International Journal of Digital Earth*, 12(5):594–613, May 2019. doi:10.1080/17538947.2018.1470688.

27 Simon Scheider, Rogier Meerlo, Vedran Kasalica, and Anna-Lena Lamprecht. Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, 20(1):167–201, June 2020. doi:10.5311/JOSIS.2020.20.555.

28 Angela Schwering. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, 12(1):5–29, February 2008. doi:10.1111/j.1467-9671.2008.01084.x.

29 Cogan Shimizu, Shirly Stephen, Rui Zhu, and colleagues. The KnowWhereGraph Ontology: A showcase. In *FOIS 2023*, volume 3637 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3637/paper46.pdf.

30 Alistair Sisson. Territory and territorial stigmatisation: On the production, consequences and contestation of spatial disrepute. *Progress in Human Geography*, 45(4):659–681, 2021. doi:10.1177/0309132520936760.

31 Nils Steinert. CoyPu Ontology. Revision: 2.3, 2023. URL: https://schema.coypu.org/global/2.3.

32 Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, and colleagues. Prithvi-EO-2.0: A Versatile Multitemporal Foundation Model for Earth Observation Applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi:10.1109/TGRS.2025.3642610.

33 Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):346, September 2024. doi:10.1093/pnasnexus/pgae346.

34 Sophie Toupin. Shaping feminist artificial intelligence. *New Media Soc.*, 26(1):580–595, 2024. doi:10.1177/14614448221150776.

35 Mike Uschold and Michael Gruninger. Ontologies: principles, methods and applications. *Knowl. Eng. Rev.*, 11(2):93–136, 1996. doi:10.1017/S0269888900007797.

36 Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, and colleagues. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9, 2016. doi:10.1038/sdata.2016.18.

37 Liling Yang, Ning Chen, Jun Yue, Yidan Liu, Jiayi Ma, Pedram Ghamisi, Antonio Plaza, and Leyuan Fang. Survey of multimodal geospatial foundation models: Techniques, applications, and challenges. arXiv preprint, 2025. doi:10.48550/arXiv.2510.22964.

38 Suhong Yoo. An operational ethical framework for geoai: A prisma-based systematic review of international policy and scholarly literature. *ISPRS International Journal of Geo-Information*, 15:51, January 2026. doi:10.3390/ijgi15010051.

39 Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: exploring data and models for visual grounding on remote sensing data. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–13, 2023. doi:10.1109/TGRS.2023.3250471.

40 Rui Zhu, Cogan Shimizu, Shirly Stephen, and colleagues. The KnowWhereGraph: A Large-Scale Geo-Knowledge Graph for Interdisciplinary Knowledge Discovery and Geo-Enrichment. *Transactions in GIS*, 30(1):e70184, 2026. doi:10.1111/tgis.70184.

Conceptual Vagueness in Human and LLM Representations: Evidence from Migration Distance Categorization

Songlin Wang[1]
Department of Geography and Regional Research, University of Vienna, Austria

Krzysztof Janowicz
Department of Geography and Regional Research, University of Vienna, Austria

Ailin Benitez Cortes
Department of Geography and Regional Research, University of Vienna, Austria

Marion Borderon
Department of Geography and Regional Research, University of Vienna, Austria

Yingjing Huang
Department of Geography and Regional Research, University of Vienna, Austria

Mina Karimi
Department of Geography and Regional Research, University of Vienna, Austria

Zilong Liu
Department of Geography and Regional Research, University of Vienna, Austria

Annika Süß
Department of Geography and Regional Research, University of Vienna, Austria

Patrick Sakdapolrak
Department of Geography and Regional Research, University of Vienna, Austria

Abstract

Conceptual vagueness is a fundamental characteristic of many geographic categories. The varied definitions of "long-distance" and "short-distance" in migration studies provides a typical example. Although interpretations of these categories are inherently context-dependent and heterogeneous, they remain widely used in domain research. This study investigates the extent to which different individuals reach a consensus on conceptualizing these categories. With a migration related corpus where "long" and "short" are generally not explicitly mentioned, we elicit human and LLM judgments in classifying each article into these two categories. Both human-human and human-LLM agreements are quantified. Our results show that human agreements are only moderate, confirming substantial conceptual variability among informed readers. LLM outputs exhibit a comparable level of variability, failing to converge on a stable category boundary. This disagreement arises from multiple sources, from detection of implicit signals to their interpretations. These findings demonstrate the feasibility of using LLMs as probes to characterize conceptual vagueness in geography at scale. Returning to migration distance, we highlight that they should not be treated as universally shared categories in knowledge organization and representation. In contrast, resolving such vagueness requires richer contextual information or additional semantic signals.

2012 ACM Subject Classification Information systems → Geographic information systems; Theory of computation → Categorical semantics

Keywords and phrases qualitative categories, conceptual vagueness, human migration, large language models, agreement

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.21

[1] Corresponding author

© Songlin Wang, Krzysztof Janowicz, Ailin Benitez Cortes, Marion Borderon, Yingjing Huang, Mina Karimi, Zilong Liu, Annika Süß, and Patrick Sakdapolrak;
licensed under Creative Commons License CC-BY 4.0

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 21; pp. 21:1–21:9

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Category Short Paper

Supplementary Material *Software (Source Code)*: `https://github.com/SonglinWang1111/Long-Short-Distance-Conceptualization` [17]
archived at `swh:1:dir:68cffc055f1b2bcd60c8893eb122fb83588c7fcf`

1 Introduction

Philosophical investigations has long recognized that many concepts and categories do not admit necessary and sufficient definitions. Such vagueness is often regarded as arising from intrinsic properties of human cognition [1]. One influential account is Wittgenstein's notion of "family resemblance", arguing that categories may lack fixed defining features, while features of diverse categories remain overlapping similarities [19]. Similarly, "prototype theory" further claims that categories exhibit graded membership centered around prototypical instances, with other entities organized progressively further, and their borderlines are generally vague [12]. In GIScience, many categories routinely used to represent spatial phenomena also lack precise and universally agreed-upon conceptualizations [9, 14]. Such categories may challenge semantic interoperability, leading to inconsistencies in spatial knowledge organization and representation, information retrieval, and policymaking [2, 6].

This vagueness is also evident in the distinction between "long-distance" and "short-distance" human migration. Since early work such as Ravenstein's laws of migration [11], these terms have been extensively used to describe and analyze mobility patterns. Similar to many other vague categories, their conceptualizations are affected not only by physical magnitudes, but also by cognitive factors such as regional background, culture, society, required time, and accessibility [5, 10, 13, 16]. Consequently, researchers have attempted to investigate how these categories are conceptualized; however, they generally sought to identify optimal metric thresholds [18]. Yet, a more fundamental question remains under-explored: to what extent people actually converge in applying these categories? Answering this question should be useful to uncover the underlying logic that triggers people's decision making.

Recent developments in large language models (LLMs) [3] offer a new methodological opportunity to dive into this research gap. Trained on massive corpora, LLMs encode patterns of language use, therefore, they can be used as statistical mirrors that reflect patterns present in human-generated discourse [15]. By aggregating LLM responses across repeated sampling, it is possible to reflect how concepts are generally represented. Besides, they have demonstrated the ability to extract, summarize, and reason over implicit knowledge in unstructured texts with limited examples of the expected input-output pairs without further training [20]. This makes LLMs promising tools for simulating human understanding of migration types through learning from recurring patterns in given examples to further categorize human migration.

In this study, we investigate how humans and LLMs categorize migration distance when it is described in text (i.e., not as numeric distance values) where, in general, neither "long-distance" nor "short-distance" is explicitly mentioned by authors. We prepare a corpus of migration-related articles, and ask human raters to classify each of them. In parallel, we prompt LLMs using few-shot examples to perform the same task. We then quantify agreements to assess consistency and variability across groups. We emphasize that neither humans nor LLMs are treated as sources of ground truth, but both are only used to reflect disagreement patterns. Meanwhile, we also emphasize that demonstrating migration categories are vague, which is a fact long recognized, is also not our purpose. Rather, we ask a different question: To what extent do humans and LLMs converge with each other when navigating through conceptually vague migration distance categories?

2 Methods

2.1 Data

This study is built upon a systematic literature review in the migration domain, which identified 53 empirical and case-study papers on "climate migration". As part of that review process, the authors initially categorized each article with respect to migration distance type (i.e., "long-distance", "short-distance", "both", or "neither"). We use the reviewed papers as our corpus and the authors' initial categorization to design the prompts and compute agreements. To ensure computational feasibility while preserve diverse contexts, we sample 33 articles from the corpus in which migration distance categories are generally not explicitly mentioned. Three of them are randomly selected as fixed in-context examples, and the remaining 30 articles are used for comparison.

2.2 Preprocessing

Before the analysis, each article is converted from PDF to structured text stored as Python dictionaries, to ensure consistent input across models, as direct PDF ingestion is not uniformly supported. Text extraction is performed using `scipdf_parser`[2]. Figures, tables, and maps are excluded in this conversion, so the categorization relies solely on textual information. We acknowledge that this may introduce slight bias because such visual elements are accessible to humans, although they rarely present explicit distance information.

2.3 Experiment Design

To assess inter-rater agreement between humans, three additional domain researchers independently categorize the selected articles without the assistance of LLMs. In parallel, we query three different LLMs under multiple settings summarized in Table 1. Both human raters and LLMs assign each article to one of four mutually exclusive categories: "long-distance migration", "short-distance migration", "both", or "neither". This scheme matches the original review and the prevailing practice in migration studies.

Table 1 Experimental settings.

Feature	Setting(s)	Rationale
Language models	GPT-4.1-mini; DeepSeek-Chat-V3.2; Gemini-2.5-Flash	Exclude model-specific influence
Prompting approaches	0-shot; 1-shot; 3-shot	Elicit implicit decision patterns
Temperatures	$T = 0;\ 0.4;\ 0.8;\ 1.2;\ 1.6$	Higher T may surface less dominant signals
Iteration	10	Mitigate the influence of stochastic variation or hallucination

Few-shot prompting approaches is used to test whether recurring patterns associated with human categorizations can be detected and reproduced by LLMs. The three fixed examples with their initial labels are provided as paired demonstrations. For the 1-shot condition, one fixed example of those three is selected to maintain coherency. Listing 1 presents the core instruction used across models.

[2] `https://github.com/titipata/scipdf_parser`

Listing 1 Core instruction in the prompt.

```
Please classify the article based on the type of human migration it
addresses by assigning one of the following codes:
'1': It focuses on or observes evidence of long-distance human
migration.
'2': It focuses on or observes evidence of short-distance human
migration.
'3': It focuses on or observes evidence of both long-distance and
short-distance human migration.
'0': It does not mention anything about human migration distance.
Do not infer any missing information. Output only the single
character without explanations, opinions or meta-comments.
```

2.4 Agreement Computation

We quantify agreement at two levels: (1) inter-rater agreements between human raters, and (2) human-model agreements between the initial reviewers and each LLM. Consistent with Section 1, neither humans nor LLMs are treated as sources of ground truth. Agreement therefore serves as a measure of consistency in categorization, not predictive correctness. For any two agents a and b, the simple agreement is defined in Equation 1, where N is the number of articles, I is the number of iterations (for human-human agreements, $I = 1$), $y_i^{(a)}$ and $y_i^{(b)}$ are the labels assigned to the same article at the same iteration by agents a and b, respectively, and $\mathbb{I}()$ is the indicator function, which equals to 1 when the two categorizations are identical and 0 otherwise.

$$\text{Agreement}(a, b) = \frac{1}{N \times I} \sum_{i=1}^{N \times I} \mathbb{I}(y_i^{(a)} = y_i^{(b)}) \tag{1}$$

For agreements between human raters, we further compute Cohen's kappa coefficient [4] which adjusts for chance agreement. Equation 2 illustrates the calculation of κ, where p_o is the observed relative agreement between raters, and p_e is the hypothetical probability of chance agreement.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{2}$$

Simple agreement ranges from 0 to 1, while Cohen's κ ranges from -1 to 1. Higher values indicate strong consistency in how category boundaries are applied.

For LLM-generated classifications, since each configuration is repeated by 10 times, aggregated agreements are reported. Given four possible categories, uniform random assignment yields an expected agreement of 25%, which serves as a baseline.

To examine the structure of disagreement, we also compute multiclass confusion matrices. These matrices help distinguish structured disagreement patterns from degenerate response distributions (e.g., overuse of a single category) and support the interpretation of agreement values as empirical manifestations of conceptual vagueness rather than response collapse.

3 Results

Table 2 summarizes the pairwise agreements and Cohen's κ coefficients among the four human raters. Pairwise agreements range from 0.40 to 0.70, while Cohen's κ ranges from 0.19 to 0.58. Although all four participants are researchers familiar with human migration, none of the pairwise comparisons reaches near-perfect agreement. These results suggest that migration distance categories are not applied according to universally shared criteria, even among domain experts. Rather than being uniformly distributed, disagreements are concentrated in a specific subset of articles, indicating that conceptual ambiguity is context-dependent. Such disagreements arise from multiple recurring forms of conceptual vagueness rather than from a single source. First, regarding several articles, some human raters consider certain linguistic signals to be sufficient in representing migration distance, while others do not think so. The second form concerns the uncertainty about where the conceptual boundary between migration distance categories should be drawn. In such articles, the contextual evidence remains compatible with multiple interpretations. Factors such as geographic scale, regional background, migration purpose, accessibility, etc. lead different raters to classify the same migration event differently.

Table 2 Pairwise agreements and Cohen's κ coefficients among the four human raters.

Human pair	Agreement	Cohen's κ
R1-R2	0.70	0.58
R1-R3	0.53	0.29
R1-R4	0.40	0.19
R2-R3	0.47	0.27
R2-R4	0.47	0.27
R1-R3	0.50	0.33

We next compare human categorizations with classifications generated by LLMs under different prompting strategies and temperatures. Figure 1 presents the agreements between each model and the original reviewers, whose categorizations are also used as demonstrations for few-shot prompting. As this figure presents, across the three models, human-model agreements remain within a range comparable to the agreements observed among human raters. In particular, GPT-4.1-mini consistently achieves agreement values similar to the human inter-rater range. These results indicate that these LLMs reproduce patterns of categorization variability that resemble those observed among human experts.

This figure also illustrates that agreement patterns remain relatively stable across temperature settings, with only minor fluctuations. Increasing sampling diversity therefore produces little influence on the final categorizations, suggesting that the observed disagreements cannot be explained primarily by stochastic generation. Instead, the variability appears to originate from the underlying conceptual ambiguity of the categorization task itself. Prompting strategy exhibits a more noticeable influence. Compared with zero-shot prompting, one-shot prompting frequently reduces agreement, whereas three-shot prompting generally restores agreement to levels comparable to or higher than the zero-shot baseline. This observation suggests that isolated demonstrations may encode localized decision patterns that do not generalize reliably, whereas multiple demonstrations provide more stable guidance regarding the implicit regularities underlying human categorization.

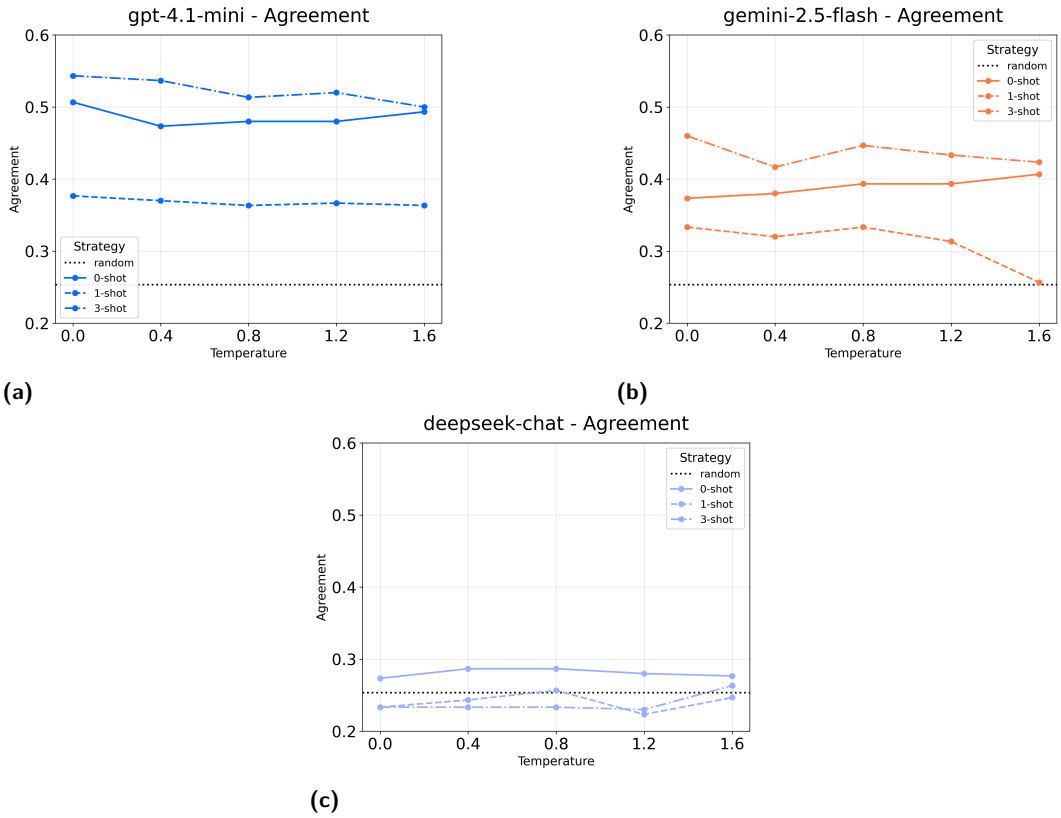


Figure 1 Human-model agreements under different temperatures and prompting strategies: (a) Human-GPT; (b) Human-Gemini; (c) Human-DeepSeek.

The qualitative observations of disagreement patterns among human raters are further supported by the confusion matrix shown in Table 3, which compares the classifications produced by GPT-4.1-mini with those of the original reviewers. Similarly, two dominant disagreement patterns can be observed. First, a considerable proportion of articles labeled as "neither" by human reviewers are classified into one of the other three migration distance categories by GPT-4.1-mini, while articles labeled as "both" by GPT are frequently not classified as this category by human reviewers. This finding indicates that agents sometimes disagree on whether sufficient evidence exists to invoke migration distance categories. Second, disagreements also occur among "long distance" and "short distance", regarding boundary of categories. Taken together, these patterns further support the interpretation that disagreement reflects structured conceptual variability rather than random prediction errors or degenerate response distributions.

Table 3 Confusion matrix comparing classifications of the initial reviewers and GPT-4.1-mini.

Human \ Model	Neither	Long distance	Short distance	Both
neither	65	130	361	344
long distance	0	251	22	177
short distance	163	233	1193	511
both	108	120	265	557

4 Discussion and Conclusion

This study moves beyond the traditional pursuit of fixed thresholds for distinguishing "long-distance" and "short-distance" migration by examining how consistently these categories are actually applied. Rather than asking whether migration distance categories are vague, which is a fact long recognized in philosophy, cognitive science, and GIScience, we investigate the extent to which humans and LLMs converge when interpreting them from textual descriptions. Across both human-human and human-LLM comparisons, moderate agreements are observed, indicating that these categories are not governed by universally shared conceptual criteria. Moreover, our observation reveals that disagreement originates from multiple forms of vagueness, regarding both the applicability of linguistic cues and the conceptual boundary between categories. These findings have implications beyond migration studies. In geographic knowledge representation, categories are often considered to possess stable and universally shared meanings. However, our results suggest that the semantics associated with "migration distance" remain inherently context dependent. Consequently, representing such kind of categories as rigid classes or fixed ontology concepts risks introducing "artificial precision" and masking legitimate conceptual diversity. Instead, geographic knowledge representation should accommodate graded membership here, considering contextual evidence, rather than assuming a single authoritative categorization.

Our findings also contribute methodologically. Rather than evaluating LLMs as classifiers, we demonstrate their potential as computational probes for conceptual analysis. While LLMs inevitably introduce biases [7, 8], several observations support our interpretation. First, similar levels of disagreements emerge among human experts and between humans and LLMs. Second, disagreement exhibits systematic structures instead of random response patterns, as reflected by both confusion matrices and qualitative analysis. Finally, agreement remains relatively stable across different temperatures, suggesting that the observed variability cannot be attributed primarily to stochastic model behavior. Together, these observations indicate that LLMs capture meaningful regularities in human conceptualizations and therefore provide a scalable methodology for investigating conceptual vagueness in geographic discourse.

Despite these contributions, several limitations remain. First, although the corpus contains linguistically rich and diverse migration narratives, it remains relatively small. Future work should investigate whether similar agreement patterns persist across substantially larger collections. Second, our corpus consists exclusively of peer-reviewed scientific publications. The conceptual patterns identified here may not generalize directly to other forms of discourse such as news reports and social media posts. Examining conceptual vagueness across multiple discourses would provide a more comprehensive understanding of how migration distance categories are socially constructed and communicated. Finally, while we identify several recurring forms of disagreement, the contextual factors underlying these patterns remain only partially understood. The exact "traces" linked to socio-cultural or infrastructural factors should be systematically investigated in future works.

In conclusion, this study provides empirical evidence that migration distance categories remain only partially shared among both humans and LLMs. Rather than indicating failures of either human or language models, the observed variability reflects the inherently vague nature of such conceptualizations. By combining human annotation with LLM outputs, we present a scalable framework for investigating conceptual vagueness. This work encourages future research toward more context-aware knowledge representations that acknowledge, rather than eliminate, conceptual uncertainty.

References

1 Ahmed Arara and Djamal Benslimane. Towards formal ontologies requirements with multiple perspectives. In Henning Christiansen, Mohand-Saïd Hacid, Troels Andreasen, and Henrik Legind Larsen, editors, *Flexible Query Answering Systems*, volume 3055, pages 150–160. Springer Berlin Heidelberg, 2004. doi:10.1007/978-3-540-25957-2_13.

2 Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pages 1–6. ACM, 2018. doi:10.1145/3227609.3227689.

3 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. Version Number: 4. doi:10.48550/arXiv.2005.14165.

4 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960. URL: https://api.semanticscholar.org/CorpusID:15926286.

5 Marián Halás and Pavel Klapka. Revealing the structures of internal migration: A distance and a time-space behaviour perspectives. *Applied Geography*, 137:102603, 2021. doi:10.1016/j.apgeog.2021.102603.

6 Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. Geospatial semantics and linked spatiotemporal data – past, present, and future. *Semantic Web*, 3:321–332, 2012. doi:10.3233/SW-2012-0077.

7 Mina Karimi and Krzysztof Janowicz. Exploring challenges of large language models in estimating the distance. *Abstracts of the ICA*, 7:68, 2024.

8 Zilong Liu, Krzysztof Janowicz, Kitty Currier, and Meilin Shi. Measuring geographic diversity of foundation models with a natural language–based geo-guessing experiment on GPT-4. In *AGILE: GIScience Series*, volume 5, pages 1–7, 2024. doi:10.5194/agile-giss-5-38-2024.

9 David M. Mark. Toward a theoretical framework for geographic entity types. In Andrew U. Frank and Irene Campari, editors, *Spatial Information Theory A Theoretical Basis for GIS*, pages 270–283, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. doi:10.1007/3-540-57207-4_18.

10 Daniel R. Montello. The perception and cognition of environmental distance: Direct sources of information. In Stephen C. Hirtle and Andrew U. Frank, editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume 1329, pages 297–311. Springer Berlin Heidelberg, 1997. doi:10.1007/3-540-63623-4_57.

11 E. G. Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235, 1885. URL: http://www.jstor.orglstable/2979181.

12 Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973. doi:10.1016/0010-0285(73)90017-0.

13 Joseph Shingleton and Ana Basiri. How close is "close"? An analysis of the spatial characteristics of perceived proximity using large language models. In *Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025*, volume 6 of *11*, pages 1–14, 2025. doi:10.5194/agile-giss-6-11-2025.

14 Barry Smith and David Mark. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B (Planning and Design)*, 30(3):411–427, 2003.

15 Johannes Treutlein, Dami Choi, Jan Betley, Sam Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 140667–140730. Curran Associates, Inc., 2024. doi:10.52202/079017-4465.

16 Yi-fu Tuan. *Space and place: the perspective of experience*. University of Minnesota Press, 1977.

17 Songlin Wang. Human Migration Distance Conceptualization. Software, swhId: `swh:1:dir:68cffc055f1b2bcd60c8893eb122fb83588c7fcf` (visited on 2026-09-01). URL: `https://github.com/SonglinWang1111/Long-Short-Distance-Conceptualization`, `doi:10.4230/artifacts.27840`.

18 Sam Wilding, David Martin, and Graham Moon. How far is a long distance? An assessment of the issue of scale in the relationship between limiting long-term illness and long-distance migration in england and wales. *Population, Space and Place*, 24(2):e2090, 2018. e2090 PSP-16-0139.R2. `doi:10.1002/psp.2090`.

19 Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA, 1953.

20 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. `doi:10.48550/arXiv.2303.18223`.

Habitual Use of Navigation Systems Impairs Spatial Learning Depending on Navigation Strategy

Qi Ying[1]
Department of Geography, University of Zurich, Switzerland

Christopher Hilton
Institute of Psychology and Ergonomics, Technische Universität Berlin, Germany

Sara I. Fabrikant
Department of Geography, University of Zurich, Switzerland

Abstract

Previous research is inconsistent on how the habitual use of navigation systems affects spatial learning of pedestrian navigators. The present study investigates whether individual differences in navigation strategy choices moderate this. Participants (N = 45) completed questionnaires assessing their navigation strategies and their reliance on pedestrian navigation systems. Spatial learning was evaluated through a real-world navigation experiment conducted in an initially unfamiliar urban environment. Participants repeatedly navigated two routes across three separate days either assisted by a mobile map or by following an experimenter. Different types of spatial knowledge were assessed using five tasks. Linear mixed-effects (LME) models were employed to analyze the relationships between measures. We found that greater habitual use of navigation systems was associated with worse route knowledge acquisition, but only for participants who relied on a map-based navigation strategy. Greater habitual use of navigation systems was also linked to worse survey knowledge acquisition, but only for mobile map-assisted navigators. This new empirical evidence contributes to a more nuanced understanding of how habitual use of navigation systems may impair spatial learning of pedestrian navigators, highlighting the moderating role of individual navigation strategies and navigation conditions.

2012 ACM Subject Classification Applied computing → Psychology; Human-centered computing → Empirical studies in HCI; Applied computing → Cartography

Keywords and phrases digital navigation aids, mobile maps, wayfinding, navigation strategy, spatial knowledge

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.22

Category Short Paper

Supplementary Material *Dataset*: `https://doi.org/10.17605/OSF.IO/E6UKX`

Funding This work was supported by the Swiss National Science Foundation [grant number 10003761] and the China Scholarship Council [grant number 202206040035].

1 Introduction

The rapid development and ubiquitous use of navigation technology have sparked debates on how habitual uses of these technologies may weaken users' innate navigation abilities over the long term. Global Navigation Satellite System technology (GNSS) has revolutionized everyday navigation by taking over cognitively demanding navigation tasks from the assisted navigator. Current location is automatically displayed on the mobile map, while turn-by-turn directions are provided with the navigation system. This reduces the need of the navigator to attend to allocentric relationships of features in the traversed environment. While navigation

[1] Geographic Information Visualization & Analysis (GIVA), University of Zurich – Irchel, Winterthurstrasse 190, 8057 Zurich, Switzerland

© Qi Ying, Christopher Hilton, and Sara I. Fabrikant;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 22; pp. 22:1–22:8

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

systems greatly enhance wayfinding effectiveness, studies have suggested that reliance on these systems may come at the cost of a deeper understanding of and interactions with the traversed environment, and thus limits spatial learning [1].

Long-term use of navigation systems could result in a decline of the users' innate spatial ability. Dahmani and Bohbot reported that greater habitual use of navigation systems during a three-year period was associated with a larger decline in spatial memory assessed in a VR maze [3]. In another VR-based study, it was found that habitual use of such systems had a negative effect on spatial learning even after controlling for the tendency of navigators with lower navigation ability to rely more heavily on navigation systems [10]. Habitual use of in-car navigation systems by expert taxi drivers impaired their spatial learning, leading to increased cognitive load during navigation [13]. In a resting-state functional magnetic resonance imaging (rs-fMRI) experiment, Fajnerová et al. found that after three months of habitual navigation system use, participants showed reduced functional connectivity between the hippocampus and other brain regions involved in navigation [4], indicating that habitual use of navigation systems is linked to brain health. In contrast, in an online study involving 822 participants, Yavuz et al. found that participants' performance in a video game requiring them to memorize a map and navigate to specified locations was not associated with their habitual use of navigation systems [12]. This finding warrants particular attention, given the very large sample size of the study and a different map-use setting.

Inconsistency in previous findings suggests that other types of individual differences may exist to explain the relationship between navigation system reliance and spatial learning such as, preferred navigation strategy. Individuals who prefer a map-based strategy during navigation tend to rely on mental representations of the environment using an allocentric reference frame to orient themselves, identify destinations, and compute routes. In contrast, individuals who prefer a scene-based strategy rely primarily on their egocentric perception and the sequential memory of landmarks and actions to reproduce previously traveled routes, which more closely aligns with the spatial information communicated by the navigation system [6]. Accordingly, individuals with different navigation strategies may be differentially affected by the habitual use of such systems. This possibility has not been well demonstrated in previous research. Moreover, most of the previous studies examined spatial learning using only a single navigation-assistance type. An exception is Ishikawa's study, which found that habitual use of in-car navigation systems deteriorated spatial learning in both mobile map-assisted and paper map-assisted navigation conditions [7].

In the present study, we investigated how habitual use of GNSS-enabled navigation systems relates to individuals' spatial learning evaluated in map-guided and experimenter-guided navigation. We specifically examined whether this relationship is moderated by individuals' navigation strategy preference. We hypothesized that greater habitual use of navigation systems is negatively associated with spatial learning of individuals who prefer a map-based navigation strategy. This is because their preferred spatial learning mechanisms may be less frequently engaged and thus less well developed during daily use of turn-by-turn navigation systems (H1). We also expected this detrimental effect to be evident across both navigation conditions (H2).

2 Methods

2.1 Participants

Forty-six participants completed the experiment. One participant was excluded from the analyses due to a misunderstanding of the task instructions. The analyzed sample consisted of 45 participants (25 females, 20 males; mean age = 27 years, SD = 3.45). The study was

registered with the ethics committee of the first author's affiliated university (No. 23.12.23). All participants provided written informed consent prior to the experiment and received 60 in the local currency as compensation for their participation in the experiment.

2.2 Experimental design

The experiment followed a within-subjects design and was conducted in a real-world urban environment initially unfamiliar to participants. Each participant navigated two routes (i.e., Route A and Route B, depicted in Figure 1); one route assisted by a mobile map (map-guided condition) and the other by following an experimenter (experimenter-guided condition). The assignment of navigation conditions to routes was counterbalanced across participants; that is, half of the participants navigated Route A with the mobile map, whereas the other half navigated it by following the experimenter. Participants navigated each route in the same manner three times (i.e., three navigation sessions) on three different days within a single week, thus allowing sufficient time for spatial knowledge development. On average, participants completed the second session 44.47 hours after the first session and the third session 119.76 hours after the first session. After navigating each route, participants' spatial learning was tested with five spatial knowledge tasks.

- **Route reversal task.** Participants were asked to reverse the route from each destination back to the starting point without any navigation assistance. If a participant took a wrong turn, the experimenter called them back to the decision point so they could make a new wayfinding choice. The time taken to complete this task was recorded in seconds, reflecting participants' ability to flexibly use different types of spatial knowledge for this wayfinding task.

Once participants reached the starting point again, they were oriented facing 112° East of true North, as depicted in Figure 1, and asked to complete the remaining four tasks.

- **Landmark recognition task.** Participants were presented with photographs of 16 visible landmarks located along the route and visible during navigation, along with eight foils. They indicated whether they had seen each presented landmark during navigation.
- **Route direction task.** The same 16 landmarks were presented again to participants, and they were asked to indicate which of three possible directions they had taken after encountering each landmark (i.e., straight, left, or right turn).
- **Pointing task and distance estimation task.** Participants indicated the direction of each landmark from their current position by drawing an arrow from the center of a circle. Their current location and facing direction were depicted on the circle. They were also asked to estimate the straight-line distance to each landmark in meters. A known distance to the nearest intersection (40 m) was provided as a reference for these distance estimates.

Using the Navigation Strategy Questionnaire [2], we found that 23 participants preferred a scene-based navigation strategy while 22 preferred a map-based navigation strategy. Participants' habitual use of pedestrian navigation systems was assessed using a self-developed background questionnaire with the question "*How often do you use pedestrian navigation systems?*" Participants selected one of the following response options: "*less than once a week,*" "*1–2 times a week,*" "*3–4 times a week,*" "*5–6 times a week,*" or "*7 times or more per week.*" Responses were coded on a scale from 1 (low frequency) to 5 (high frequency) for subsequent analyses.

Figure 1 (A) The two navigated routes shared the same starting point and were similar in length (Route A: approximately 750 m; Route B: approximately 680 m) and comparable in structure, with 10 intersections and seven turns. (B) The mobile map contained a base map, the start and destination locations, the route to be followed and the navigator's current location. The labels are only included for explanation in this Figure and were not shown on the mobile map during the experiment. (C) A sample landmark photo used for the spatial knowledge test.

2.3 Procedure

Participants were welcomed to a university laboratory located near the experimental area, where they first gave informed consent and then completed various questionnaires. After the setup of the EEG and the eye tracker, participants were escorted to the starting point near the university and given instructions for the outdoor portion of the experiment. They then performed the route-following, route-reversal, and spatial knowledge tasks for the first route. The same procedure was repeated for the second route they navigated.

2.4 Data analysis

Participants' performance on the landmark recognition task, reflecting their ability to acquire landmark knowledge, was assessed using the discriminability index d prime (d' = z[hit rate] - z[false alarm rate]) based on signal detection theory [11]. Route direction recall accuracy (binary: correct or incorrect) reflected participants' ability to acquire route knowledge. Pointing and distance estimation performance were evaluated using absolute pointing error (in degrees) and the absolute distance estimation error (in meters), respectively. These two indices reflect participants' ability to acquire survey knowledge.

Outlier correction was conducted for each task. For each navigation condition and session combination, participants with extremely poor performance were identified as outliers using a 3 × standard deviation (SD) criterion and were removed from the analyses. In total, six outliers were identified for the pointing task, three for the distance estimation task, and two for the route reversal task. We fitted linear and generalized linear mixed effect (LME and GLME) models to predict spatial knowledge test performance. The models included the following fixed effects: navigation condition (using sum contrasts), session (using successive differences contrasts), navigation strategy (using sum contrasts), and the frequency of navigation system use (continuous). All models incorporated two-way interaction terms for each pair of predictors and included participant ID as a random factor with random intercepts. All analyses were conducted in R (v. 4.4.2).

3 Results

We found a significant main effect of frequency of navigation system use on pointing error. Higher frequency of navigation system use was associated with increased pointing error (β = 2.81, SE = 1.28, t = 2.20, p = .034). No significant main effects of navigation system use were observed for the other tasks in the spatial knowledge test (all $p > .05$).

Regarding the moderating effect of navigation strategy, the LME model for the route direction task revealed a significant interaction between frequency of navigation system use and preferred navigation strategy (β = 0.18, SE = 0.07, t = 2.73, p = .006). Post-hoc analyses indicated that frequency of navigation system use significantly predicted route direction recall accuracy for participants who preferred a map-based navigation strategy (β = -0.19, SE = 0.09, z = -2.12, p = .034), but not for participants who preferred a scene-based navigation strategy (β = 0.17, SE = 0.10, z = 1.77, p = .078). The contrast between these two strategy slopes was also significant (β = 0.36, SE = 0.13, z = 2.73, p = .006). Visual inspection of Figure 2A confirmed that greater habitual use of navigation systems was associated with poorer route direction recall performance, but only for map-based navigators.

Significant interactions between the frequency of navigation system use and navigation condition emerged in the LME models for the pointing task (β = 0.69, SE = 0.33, t = 2.05, p = .041) and the route reversal task (β = -3.33, SE = 1.48, t = -2.26, p = .025). As observed in Figure 2B and examined in post-hoc analyses, higher frequency of navigation system use was associated with increased pointing error in the map-guided condition (β = 3.50, SE = 1.32, z = 2.65, p = .008), but not in the experimenter-guided condition (β = 2.13, SE = 1.32, z = 1.61, p = .108). The contrast between these two condition slopes was significant (β = 1.37, SE = 0.67, z = 2.05, p = .041). For the route reversal task (Figure 2C), a different pattern was observed: the negative association between the frequency of navigation system use with route reversal performance appeared more prominent for the experimenter-guided condition. Despite a significant difference between the condition slopes (β = -6.66, SE = 2.95, t = -2.26, p = .025), the frequency of navigation system use did not significantly predict route reversal performance in either navigation condition, as examined with post-hoc analyses (all ps > .05).

4 Discussion

This study set out to investigate the inconclusive impact of habitual navigation system use on spatial learning in varying navigation assistance conditions, by examining whether this relationship might differ across wayfinders' preferred navigation strategies. We found that the habitual use of navigation systems impaired route knowledge acquisition for participants who preferred an allocentric, map-based navigation strategy, but not other forms of spatial knowledge, partially supporting Hypothesis 1. Our findings also indicated that increased habitual use of navigation systems was associated with greater pointing error, reflecting poorer survey knowledge acquisition, consistent with those reported by Ruginski et al. [10]. Contrary to our prediction that the deteriorating effect would be universal (Hypothesis 2), it was only significant in the map-guided condition.

Taken together, these findings could mean that navigation systems predominantly provide and emphasize egocentric information (e.g., turn-by-turn instructions at intersections), which is relatively easy to process as it requires less cognitively demanding cognitive mechanisms [9]. Consequently, users of such systems may be less spontaneously engaged in processing allocentric map information and thus in forming cognitive maps, which require higher-level learning mechanisms [8]. Somewhat paradoxically, greater reliance on navigation systems may

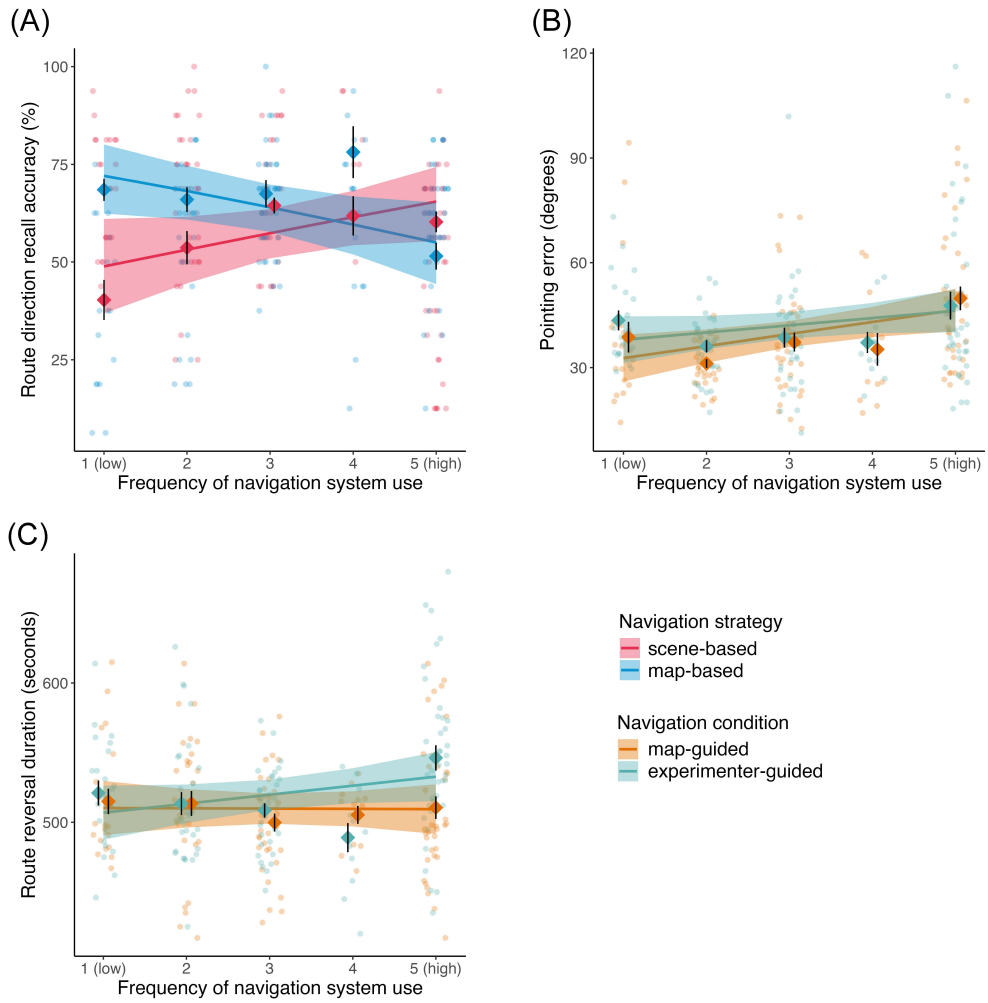


Figure 2 (A) The frequency of navigation system use was negatively associated with route direction recall accuracy among participants who preferred a map-based navigation strategy, but not among those who preferred a scene-based navigation strategy. (B) The frequency of navigation system use significantly predicted pointing error only in the map-guided condition. (C) The negative effect of navigation system use frequency on route reversal performance appeared to be more pronounced in the experimenter-guided condition. However, the association between navigation system use frequency and route reversal duration was not significant in either the map-guided or experimenter-guided condition. Error bars represent standard errors, and shaded areas indicate 95% confidence intervals. Data points were jittered along the x-axis to reduce overlap.

therefore impair navigators' ability to acquire survey knowledge from both the environment and maps, leading to poorer pointing performance, particularly in the map-guided condition where maps were available. This may explain the pattern shown in Figure 2B, where the smaller pointing error in the map-guided condition compared to the experimenter-guided condition observed among participants with lower navigation system use frequency disappears for those with higher use frequency. For navigators who preferred a map-based navigation strategy, habitual use of navigation systems not only impaired their survey knowledge acquisition, but also hindered route knowledge acquisition. One possible explanation is that these individuals typically rely on allocentric processing to encode route information. However, this strategy may be underutilized during habitual navigation system use, resulting in a decline in route learning.

An alternative explanation for the findings in Figure 2B is that navigators who use navigation systems less frequently may possess higher spatial ability, making them more proficient at acquiring allocentric cues directly from maps [5], and thus better positioned to benefit from map availability in the map-guided condition. Consequently for these individuals, differences in pointing performance between the map-guided and experimenter-guided conditions may be more pronounced.

5 Conclusion and future work

This empirical outdoor navigation study conducted in an urban environment indicates that long-term reliance on GNSS-based pedestrian navigation systems negatively influences the ability in route knowledge acquisition, but only for users who employ a map-based navigation strategy. Such reliance is also associated with reduced ability to acquire survey-level direction knowledge, but only during mobile map-assisted navigation. These findings enhance our understanding of how reliance on pedestrian navigation systems can impair users' spatial learning and shed light on the moderating roles of individual navigation strategies and navigation conditions in this relationship.

One limitation of this study is the relatively small sample size, which makes it difficult to cover participants across a broader range of ages, occupations, living in different cities, etc. Furthermore, we assessed participants' reliance on navigation systems with a single, self-reported question, and only considered pedestrian navigation systems. Future research could further explore comprehensive assessments of participants' use of different types of navigation systems (e.g., in-car navigation systems) and quantifying their habits over different time periods (i.e., the last year, the last ten years, etc.). Moreover, recording usage data directly from participants' personal mobile devices (e.g., smartphones) would allow for a more accurate measurement of navigation system use and provide deeper insight into how these systems are employed in daily life.

References

1 Eran Ben-Elia. An exploratory real-world wayfinding experiment: A comparison of drivers' spatial learning with a paper map vs. turn-by-turn audiovisual route guidance. *Transportation Research Interdisciplinary Perspectives*, 9:100280, March 2021. doi:10.1016/j.trip.2020.100280.

2 Iva K. Brunec, Buddhika Bellana, Jason D. Ozubko, Vincent Man, Jessica Robin, Zhong-Xu Liu, Cheryl Grady, R. Shayna Rosenbaum, Gordon Winocur, Morgan D. Barense, and Morris Moscovitch. Multiple scales of representation along the hippocampal anteroposterior axis in humans. *Current Biology*, 28(13):2129–2135.e6, July 2018. doi:10.1016/j.cub.2018.05.016.

3 Louisa Dahmani and Véronique D. Bohbot. Habitual use of gps negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10(1), April 2020. doi:10.1038/s41598-020-62877-0.

4 Iveta Fajnerová, David Greguš, Jaroslav Hlinka, Tereza Nekovářová, Antonín Škoch, Tomáš Zítka, Jan Romportl, Eva Žáčková, and Jiří Horáček. Could prolonged usage of gps navigation implemented in augmented reality smart glasses affect hippocampal functional connectivity? *BioMed Research International*, 2018:1–10, June 2018. doi:10.1155/2018/2716134.

5 Chuanxiuyue He and Mary Hegarty. How anxiety and growth mindset are linked to navigation ability: Impacts of exploration and gps use. *Journal of Environmental Psychology*, 71:101475, October 2020. doi:10.1016/j.jenvp.2020.101475.

6 Mary Hegarty, Chuanxiuyue He, Alexander P. Boone, Shuying Yu, Emily G. Jacobs, and Elizabeth R. Chrastil. Understanding differences in wayfinding strategies. *Topics in Cognitive Science*, 15(1):102–119, January 2022. doi:10.1111/tops.12592.

7 Toru Ishikawa. Satellite navigation and geospatial awareness: Long-term effects of using navigation tools on wayfinding and spatial orientation. *The Professional Geographer*, 71:197–209, 2019. doi:10.1080/00330124.2018.1479970.

8 Tobias Meilinger, Julia Frankenstein, Katsumi Watanabe, Heinrich H. Bülthoff, and Christoph Hölscher. Reference frames in learning from maps and navigation. *Psychological Research*, 79(6):1000–1008, November 2014. doi:10.1007/s00426-014-0629-6.

9 Renato Orti, Tina Iachini, Eleonora D'Agostino, Francesco Ruotolo, and Gennaro Ruggiero. Cognitive load in switching between egocentric and allocentric spatial frames of reference: a pupillometry study. *Scientific Reports*, 15(1), July 2025. doi:10.1038/s41598-025-11122-7.

10 Ian T. Ruginski, Sarah H. Creem-Regehr, Jeanine K. Stefanucci, and Elizabeth Cashdan. Gps use negatively affects environmental learning through spatial transformation abilities. *Journal of Environmental Psychology*, 64:12–20, August 2019. doi:10.1016/j.jenvp.2019.05.001.

11 Wilson P. Tanner and John A. Swets. A decision-making theory of visual detection. *Psychological Review*, 61:401–409, November 1954. doi:10.1037/H0058700.

12 Emre Yavuz, Chuanxiuyue He, Christoffer J. Gahnstrom, Sarah Goodroe, Antoine Coutrot, Michael Hornberger, Mary Hegarty, and Hugo J. Spiers. Video gaming, but not reliance on gps, is associated with spatial navigation performance. *Journal of Environmental Psychology*, 96:102296, June 2024. doi:10.1016/j.jenvp.2024.102296.

13 Qi Ying, Weihua Dong, and Sara Irina Fabrikant. How do in-car navigation aids impair expert navigators' spatial learning ability? *Annals of the American Association of Geographers*, 114:1483–1504, 2024. doi:10.1080/24694452.2024.2356858.

Capturing Environmental Scale Spatial Cognition in Applied Settings: A Field Based Pilot with Firefighters

Chelsie McWhorter[1]
Geospatial Cognition Lab, Oklahoma State University, Stillwater, OK, USA
University of California Santa Barbara, CA, USA

Kathy Baylis
Department of Geography, University of California Santa Barbara, CA, USA

Mary Hegarty
Department of Psychological & Brain Sciences, University of California Santa Barbara, CA, USA

Abstract

Understanding how spatial knowledge is organized and used across real-world tasks remains an important challenge in spatial cognition research. This paper presents a pilot test of SPaRKS, a framework developed to capture cognitive map accuracy, route decision making, verbalized navigation strategies, GPS use, and navigational value across a dynamic set of cartographic tasks and questionnaires. Using firefighters as an applied pilot group (n=24), we examine whether the tool can capture both route attribute tradeoffs and broader variation in spatial knowledge across individuals and groups. Results indicate that route choice was shaped primarily by distance and steep grades, suggesting that participants did not rely on shortest-path logic alone, but instead weighed practical roadway characteristics when making routing decisions. The pilot also identified meaningful variation across multiple spatial measures, indicating that SPaRKS can move beyond isolated task performance to support the development of integrated spatial profiles. Taken together, this preliminary study illustrates that SPaRKS is a viable framework for examining environmental scale spatial knowledge in applied settings. More broadly, the paper demonstrates the value of a multimodal approach for studying how spatial knowledge is organized and used across route decisions, cognitive map tasks, and self-reported navigation tendencies.

2012 ACM Subject Classification Information systems → Geographic information systems; Applied computing → Operations research; Human-centered computing → Empirical studies in HCI

Keywords and phrases Cognitive maps, navigation, route decision making, operational data collection, tool creation

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.23

Category Short Paper

Acknowledgements The authors wish to thank the firefighters who pilot tested SPaRKS, we appreciate their time, hospitality, and commitment to advancing research.

1 Introduction

Spatial abilities help us do everything from locating apps on our smartphones, to finding our way around a new building or city. These fundamental cognitive tools include our ability to find our way, understand perspective and scale, develop cognitive maps [12], and estimate distances [8] (among others). As technology accessibility and computing power increase, we have the opportunity to study these abilities in more detail than ever before.

[1] Corresponding author

© Chelsie McWhorter, Kathy Baylis, and Mary Hegarty;
licensed under Creative Commons License CC-BY 4.0

Much spatial cognition research has relied on controlled tasks and non-operational samples, and often at figural or vista scale [19]. While laboratory settings provide important and controlled empirical work on topics such as spatial orientation, perspective taking, and navigation [13, 8, 23, 20], complementary field based approaches can help examine how people perform these tasks in everyday contexts and at environmental scale [12, 5, 1]. However, capturing environmental spatial knowledge in applied settings can be problematic. Nonetheless, measuring a variety of spatial abilities and wayfinding inclinations outside of the laboratory, in real world settings, allows for environmental scale measurement and expands our research participants to a larger group of individuals, since they are not required to attend our laboratories. The purpose of this study is to develop and pilot a novel instrument for extracting and analyzing the structure of cognitive spatial knowledge in field based settings.

While many tools exist to measure spatial ability in laboratory settings, few methods capture the complexity of real-world navigation, environmental familiarity, and decision making in a professional high-stakes setting. To overcome these challenges, a multipart instrument was designed and deployed on tablet computers. The Spatial and Representational Knowledge Survey (SPaRKS) utilizes a standard tablet with internet connectivity to capture these data. The SPaRKS instrument measures spatial thinking through 3 main tasks: (1) a pin-drop mapping exercise to elicit cognitive representations of spatial layouts; (2) a turn-by-turn direction task to investigate participant spatial strategies; and (3) a discrete choice task to assess the physical attributes that are given priority in route selection (e.g. selecting routes based on if they contain steep gradients, have a tight turn radius, or are the shortest distance). In addition to these tasks, the instrument also collects self-report data on sense of direction, the value placed on navigation, GPS use, and demographics. Taken together these tasks demonstrate how this integrated instrument can reveal spatial reasoning processes that are often difficult to measure, especially in applied settings. To demonstrate the capability of SPaRKS, a pilot study was conducted with firefighters in a municipal fire department in Southern California.

2 Background

Studying environmental scale spatial cognition broadens our understanding beyond traditional research populations and includes those who make spatial decisions in high-stakes environments, such as firefighters. Further, environmental scale spatial cognition helps us connect laboratory findings to how spatial cognition is expressed in navigation relevant environments. Studies of individual differences in navigation and spatial cognition reveal that some people exhibit a stronger ability to develop knowledge of places than others [12, 2], and that strategies for doing so also differ [7, 20].

Many real-world navigation problems occur at environmental scale, where decisions happen in relation to routes, constraints, and local context. Environmental scale decisions, like many spatial cognition processes, are influenced by both more objective spatial abilities as well as more subjective spatial inclinations. Capturing objective spatial ability has been done using a variety of instruments, the Money Road Map test [18], the Spatial Orientation Test [10], Silcton [23], and Sea Hero Quest [21], among others.

Subjective measures of spatial cognition are also valuable in understanding spatial behaviors and decision making. Individual self-assessment of sense of direction correlates highly with objective spatial performance [9]. Frequent use of navigation systems has been linked to weaker spatial performance, including lower self-reported sense of direction and greater reliance on route-based strategies [11, 17]. Similarly, a recent meta-analysis found

that GPS use was negatively correlated with both sense of direction and environmental knowledge [17]. However, GPS use is dynamic and nuanced, some may use it for gathering information, such as checking traffic, locating a nearby service, or planning a route, while others may use it for turn-by-turn directions [22]. More research is needed to understand the directional relationship of many of these constructs, nonetheless they provide insight into how spatial decisions are executed.

Previous research has found that route decision making does not always result in taking the shortest path [3], the clearly marked path, or the path with rich details [24], which can lead to delays that are at best annoyances and at worst detrimental to life and health. Refining our understanding of how route decisions are made in real-world contexts can help emergency managers better plan evacuation routes, ease traffic congestion, defend pedestrian safety, and allow emergency responders to plan responses more effectively.

3 Methods

SPaRKS is comprised of objective measurement tasks, subjective wayfinding inclination questionnaires, and demographic data collection. Participants are presented with a customized map of an environment that is familiar to them which contains building footprints without names, streets but no street names, and at least 2 named landmarks to help them situate the scale and layout of the map. Next, they are presented with an address and asked place a digital map-pin onto the map where they think that address is located. Correctness is determined by both context and a GIS calculated buffer zone. The address selection task has 4 individual addresses, each presented one at a time.

Next, participants are presented with another set of addresses, one at a time, and asked to write out how they would route to that address given a predetermined starting point. The written directions portion contains 6 addresses. These written routes are transcribed onto maps in a GIS and a similarity analysis is run. Further, the routes are also qualitatively analyzed using content analysis [6] to explore how their spatial language is related to their strategy and performance [14].

Finally, participants are presented with one screen that contains 3 maps panels, each panel contains a different route between a single starting point and a destination. The maps featured local streets and street names familiar to the participants but excluded details such as traffic or topography. This design explicitly required participants to synthesize the visual map prompt with their existing internal cognitive maps of the response district. Participants are asked to choose which route they would take if they were trying to get to the destination the fastest. Participants are presented with 6 total route-choice questions. These data are compiled and analyzed following discrete choice analysis via multinomial logit regression, with consideration of physical route characteristics (length of the route, number of turns, and steepness of the road grade) as well as the source of the route (Figure 1). Note that, firefighter navigation training in the United States is non-uniform; departments predominantly rely on informal methods like map memorization, street-identification games, and district driving [16]. Consequently, the map tasks used in this pilot study (pin-drop, turn-by-turn description, and route selection) are not established methods for how firefighters typically learn their response areas.

The routes and destinations used in SPaRKS were chosen to represent different levels of difficulty. To achieve varying levels of difficulty, the Route Choice Scenario Builder was developed in RStudio using the Shiny package [4] to assist with categorizing difficulty.

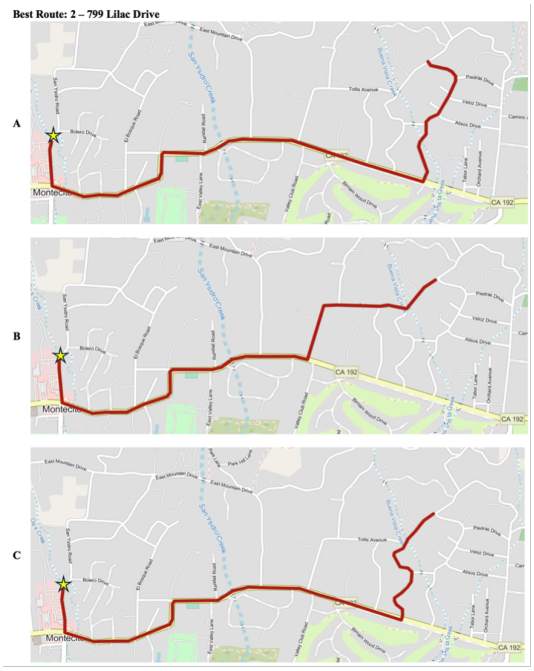


Figure 1 An example of one of the 3 maps panels presented to particpants for the route choice questions.

SPaRKS also captures self-report measures of spatial cognition including the Santa Barbara Sense of Direction scale (SBSOD) [9], value placed on navigation, and the GPS use questionnaire [22]. Participants are also given a demographic questionnaire to capture their age, gender, education level, and other applicable details.

4 Procedure

To assess the functionality of SPaRKS, it was pilot tested with a municipal fire department in Southern California. A total of 24 male firefighters (average age 42 years) participated in the study while on duty. SPaRKS was delivered on iPad tablets with internet connectivity at their respective firehouses. The researcher introduced the topic, explained the procedure and asked the participants to work alone on their survey. Participants were instructed to ask questions if they were not clear on the procedure, and the researcher remained in the room with the participants during data collection. At the end of each set of data collection, the firefighters provided feedback on the survey and were given the opportunity to ask questions about the project. They were instructed to not give any tips or other survey information to the other firefighters that were not on shift at the present data collection visit. In total data collection was completed over 5 visits to the firehouses.

5 Results

SPaRKS identified patterns in route selection that illustrate how participants prioritize route characteristics such as grade, number of turns, and length, features that would be difficult to capture through surveys alone. Each participant was presented with 6 destinations for the route choice task, which was designed to elucidate spatial decision making and identify preferable route attributes. The focus here is on the route-choice findings because they most directly showcase SPaRKS ability to capture operational spatial decision making.

Multinomial logit modeling on physical route characteristics identified that route length and the presence of steep grades were significant predictors of route choice. For each 1-mile increase in route length, the odds of that route being selected decreased by 65%, and if the route contained a steep grade the odds of selecting it dropped by 94%. Number of turns was not a statistically significant predictor (Table 1). Analysis of utility tradeoffs for how distance influenced choice, specifically how much participants were willing to travel to reduce the number of turns or avoid steep grades, found that on average, participants were willing to add 2.69 miles onto their route if it avoided steep grades, while number of turns did not reach significance. (Table 1).

Table 1 Route Attributes; multinomial logit model.

Predictor	β	Robust SE	Odds Ratio	95% CI	p
Distance (miles)	-1.05	0.32	0.35	[0.19, 0.66]	0.001
Number of Turns	0.10	0.10	1.10	[0.90, 1.34]	0.34
Steep grade	-2.83	0.82	0.06	[0.01, 0.29]	$< .001$

Note. β ***= coefficient; OR = odds ratio***

Overall, these findings suggest that firefighters did not simply rely on shortest-path logic when selecting routes. Instead, route choice reflected practical tradeoffs tied to roadway suitability, with distance remaining important, but steep grades also shaping what routes were seen as preferable. Although source was also related to route choice, these differences appear to be driven primarily by the physical characteristics of the routes themselves.

SPaRKS was able to capture individual variation in cognitive map accuracy, route selection, verbalized navigation strategies, GPS use, and the value individuals place on navigation. Findings from this pilot test suggest that SPaRKS goes beyond isolated task performance and provides data that can be compiled into an integrated spatial profile, offering a way to capture multidimensional variation in spatial knowledge. An integrated spatial profile shifts the focus away from a singular spatial knowledge type, or simply asking who is a "good navigator," and instead toward a more complete understanding of what kinds of spatial knowledge and strategies are present within a person or group.

To explore the creation of an integrated spatial profile, performance was compared across the components of SPaRKS [15]. Firefighters holding the rank of Engineer (driver/operator) performed stronger than other ranks, particularly in the pin-dropping task and on the GPS use questionnaire, while those with more experience also showed less pin-drop error, lower GPS use, and stronger navigational value scores. Taken together, these findings suggest that SPaRKS is sensitive not only to route decision making, but also to broader differences in how spatial knowledge is organized and used across individuals.

6 Discussion

Within the fire service, these data offer immediate insights into who needs more training, assess whether their current navigation training is effective, and isolates firefighters for promotion into the engineer (driver/operator) position. During post data collection discussions with the firefighters who participated in the pilot, many expressed interest to know how well they performed across the tasks. Further, the department leadership requested aggregate data to systematically rectify training deficiencies and refine processes. Beyond the scope of the

fire service, SPaRKS could be a useful tool for other emergency responders such as police, emergency medical providers, as well as search and rescue teams, where high-stakes and time sensitive response is critical to the outcome.

SPaRKS also shows utility outside emergency responders, with application into the evaluation of groups such as novice or elderly drivers, commercial delivery couriers, drone pilots, or orienteering competitors. These groups could provide broader insight into the development and decline of spatial abilities. Such research would clarify how specific operational demands intersect with spatial cognition, whether managing the 2D-to-3D perspective shifts required of drone operators, the race pressure experienced by competitive athletes, or the temporal pressures faced by emergency responder professionals trained to handle time-sensitive and high-stakes responses.

This study provides evidence that SPaRKS can extend and complement laboratory based spatial cognition research toward a more applied understanding of how people know, reason through, and move across real environments. Future work should utilize more participants, and include a more heterogeneous sample with regards to age, gender, and profession.

7 Conclusion

This preliminary study illustrates that SPaRKS is a useful tool for capturing field based spatial cognition across a variety of tasks. Rather than isolating spatial knowledge through a single task, SPaRKS was able to capture variation in cognitive map accuracy, route decision making, verbalized navigation strategies, GPS use, and navigational value. Importantly, these patterns were not random; they varied in meaningful ways with familiarity, experience, and operational role, suggesting that SPaRKS is sensitive to real-world differences in how spatial knowledge is organized and used. The route choice findings further indicate that field based approaches can capture spatial decision making in context, including the tradeoffs individuals make between distance, roadway characteristics, and operational suitability.

Overall, this pilot study provides initial evidence that SPaRKS is a viable framework for exploring environmental scale spatial cognition in applied settings. By integrating multiple forms of spatial data, SPaRKS offers a way to connect controlled spatial cognition research to the applied context in which navigation and route decisions unfold.

References

1 C. Bae, D. Montello, and M. Hegarty. Wayfinding in pairs: Comparing the planning and navigation performance of dyads and individuals in a real-world environment. *Cognitive Research: Principles and Implications*, 9(1):40, 2024. doi:10.1186/s41235-024-00563-9.

2 A. P. Boone, B. Maghen, and M. Hegarty. Instructions matter: Individual differences in navigation strategy and ability. *Memory & Cognition*, 47(7):1401–1414, 2019. doi:10.3758/s13421-019-00941-5.

3 M. L. Brachman, R. Church, B. Adams, and D. Bassett. Wayfinding during a wildfire evacuation. *Disaster Prevention and Management*, 29(3):249–265, 2020. doi:10.1108/DPM-07-2019-0216.

4 W. Chang, J. Cheng, J. Allaire, et al. *shiny: Web Application Framework for R*, 2024. R package version 1.10.0. URL: https://CRAN.R-project.org/package=shiny.

5 E.-M. Griesbauer, E. Manley, J. M. Wiener, and H. J. Spiers. London taxi drivers: A review of neurocognitive studies and an exploration of how they build their cognitive map of London. *Hippocampus*, 32(1):3–20, 2022. doi:10.1002/hipo.23395.

6 I. Hay and M. Cope, editors. *Qualitative research methods in human geography*. Oxford University Press, 5th edition, 2021.

7 M. Hegarty, C. He, A. P. Boone, et al. Understanding differences in wayfinding strategies. *Topics in Cognitive Science*, 15(1):102–119, 2023. doi:10.1111/tops.12592.

8 M. Hegarty, D. R. Montello, A. E. Richardson, et al. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2):151–176, 2006. doi:10.1016/j.intell.2005.09.005.

9 M. Hegarty, A. E. Richardson, D. R. Montello, et al. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002. doi:10.1016/S0160-2896(02)00116-2.

10 M. Hegarty and D. Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2):175–191, 2004. doi:10.1016/j.intell.2003.12.001.

11 T. Ishikawa. Satellite navigation and geospatial awareness: Long-term effects of using navigation tools on wayfinding and spatial orientation. *The Professional Geographer*, 71(2):197–209, 2019. doi:10.1080/00330124.2018.1479970.

12 T. Ishikawa and D. R. Montello. Spatial knowledge acquisition from direct experience in the environment. *Cognitive Psychology*, 52(2):93–129, 2006. doi:10.1016/j.cogpsych.2005.08.003.

13 M. Kozhevnikov and M. Hegarty. A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29(5):745–756, 2001. doi:10.3758/BF03200477.

14 T. P. McNamara, J. Sluzenski, and B. Rump. Human spatial memory and navigation. In *Learning and Memory: A Comprehensive Reference*, pages 157–178. Elsevier, 2008. doi:10.1016/B978-012370509-9.00176-5.

15 C. McWhorter. *Navigation & Spatial Cognition in the US Fire Service*. PhD thesis, University of California, Santa Barbara, 2025. URL: https://escholarship.org/uc/item/4586s2vj.

16 C. McWhorter, M. Hegarty, K. Baylis, et al. Mapping the response: A survey of municipal firefighter navigation training and practices in the United States. *International Journal of Disaster Risk Reduction*, 123:105446, 2025. doi:10.1016/j.ijdrr.2025.105446.

17 L. Miola, V. Muffato, E. Sella, et al. GPS use and navigation ability: A systematic review and meta-analysis. *Journal of Environmental Psychology*, page 102417, 2024. doi:10.1016/j.jenvp.2024.102417.

18 J. Money, D. Alexander, and H. T. Walker. *A Standardized Road Map Test of Direction Sense*. Johns Hopkins Press, 1967.

19 D. R. Montello. Scale and multiple psychologies of space. In *Spatial Information Theory A Theoretical Basis for GIS*, volume 716, pages 312–321. Springer, 1993. doi:10.1007/3-540-57207-4_21.

20 V. Muffato, L. Miola, M. Pellegrini, et al. Investigating the different domains of environmental knowledge acquired from virtual navigation. *Cognitive Research: Principles and Implications*, 8(1):50, 2023. doi:10.1186/s41235-023-00506-w.

21 H. J. Spiers, A. Coutrot, and M. Hornberger. Explaining world-wide variation in navigation ability from millions of people: Citizen science project Sea Hero Quest. *Topics in Cognitive Science*, 15(1):120–138, 2023. doi:10.1111/tops.12590.

22 A. Topete, C. He, J. Protzko, et al. How is GPS used? understanding navigation system use and its relation to spatial ability. *Cognitive Research: Principles and Implications*, 9(1):16, 2024. doi:10.1186/s41235-024-00545-x.

23 S. M. Weisberg, V. R. Schinazi, N. S. Newcombe, et al. Variations in cognitive maps: Understanding individual differences in navigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):669–682, 2014. doi:10.1037/a0035261.

24 J. Wilkening and S. I. Fabrikant. How do decision time and realism affect map-based decision making? In *Spatial Information Theory*, pages 1–19. Springer, 2011. doi:10.1007/978-3-642-23196-4_1.

Context-Sensitive Spatial Deixis Through the Lens of Mental Models

Adwitiya Mandal
Department of Computing Science, Umeå University, Sweden

Zoe Falomir
Department of Computing Science, Umeå University, Sweden

Kai-Florian Richter
Department of Computing Science, Umeå University, Sweden

Abstract

The use of spatial deixis, such as *here* and *there*, depends not only on physical distance, but also on how speakers mentally organize the relevant space. This paper examines how contextual factors may influence and shape the use of spatial deixis in an object-selection task. The analysis is based on verbal data from a previous human-robot interaction study, in which 12 participants referred to objects in a shared cupboard environment. Using Norman's distinction of socio-cognitive systems, we interpret the observed deixis behavior as evidence for potential spatial mental models underlying participants' spatial deixis use. We find that speakers' deixis use reflects a combination of absolute, relative, mixed, temporal, and response-biased models rather than a single fixed strategy. Our findings support the view that spatial deixis is context-sensitive, shaped by the mental organization of the space and is variable across individuals, with implications for interpreting spatial deixis in human-robot interaction.

2012 ACM Subject Classification Human-centered computing → Empirical studies in HCI; Human-centered computing → HCI theory, concepts and models; Applied computing → Psychology; Computing methodologies → Cognitive robotics

Keywords and phrases Spatial deixis, here and there, mental models, human-robot interaction, object selection, spatial cognition

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.24

Category Short Paper

Supplementary Material *Dataset (Excel table)*: `https://osf.io/qh5ap/overview`

Funding *Adwitiya Mandal*: funded by the Swedish Research Council (VR) under grant 2022-03966. *Zoe Falomir*: acknowledges the funding by the Wallenberg AI, Autonomous Systems and Software Program (WASP) awarded by the Knut and Alice Wallenberg Foundation, Sweden.

1 Introduction

Referring to physical space is highly context-dependent [10], ranging from topological relationships between objects to directional spatial relations within larger environments, such as rooms, where other physical entities play a role in defining spatial actions. In all languages, demonstratives play a crucial role in discourse. In English, demonstratives are classified as "proximal" (*this*, *these* and *here*) and "distal" (*that*, *those* and *there*) [1, 12]. They are closely related and are defined by the same deictic roots of joint attention [7, 5], which is considered a fundamental element of social cognition. Among them, spatial deixis (*here* and *there*), are words that depend on the spatial position of the speaker and/or listener and convey location-specific information [16, 9].

The interpretation of spatial deixis depends less on the objective metric space and more on how the speaker conceptualizes the communicative context [8]. Neuro-psychological studies suggest that such spatial distinctions in demonstratives are rooted in embodied interaction

© Adwitiya Mandal, Zoe Falomir, and Kai-Florian Richter;
licensed under Creative Commons License CC-BY 4.0

with one's own space, particularly in the distinction between one's peri-personal and extra-personal space. Peri-personal space is the region of space within a person's reach, where they can interact with objects in their immediate visuo-tactile surroundings by reaching, grasping or touching [2, 3, 13, 14]. In contrast, extra-personal space is the space beyond that where objects are only perceptually visible, but physically inaccessible. The peri-personal space is contextually dynamic. For example, it increases if the person has a tool in their hand. In this context, Skilton suggests that experimental and observational studies highlight these dynamic peri-personal spaces as the "here-space" [19]. These context sensitivity and shared attention perspectives of spatial deixis pose both linguistic and perceptual challenges for robots to understand their human collaborators in an interaction scenario.

Caldano and Coventry [4], building on discussions of the deictic content of demonstratives [19], point out that their usage depends not only on the perceptual space, but also factors, such as object properties, the hearer's position and joint attention. Extending this view, the present paper investigates the research question "how does context influence the use of the spatial deixis 'here' and 'there' (in an object selection task)?"

2 The experimental setting

The present study builds on verbal interaction data collected in a previous Human-Robot Interaction (HRI) experiment [15]. In this experiment, participants were instructed to direct the robot in an imagined scenario of moving objects from a cupboard. It did not actually involve any movement of the robot or the objects. Participants pointed to cups and bowls placed on the cupboard while facing it from the same perspective as the robot, with both remaining at fixed positions through-out the interaction. The study comprised of five different scenarios where the objects were arranged in increasing density and in different configurations, as shown in Fig 1. In each scenario, the robot asked the participant to select an object by both pointing and referring verbally to that object. Participants were instructed to use a verbal expression that at least contained the object class along with a spatial deixis, as they would deem appropriate, for example, "the cup there." These verbal instructions were recorded for all the objects across all scenarios for 12 participants. These data form the basis of the present analysis.

Figure 1 The five experimental scenarios showing the arrangement of objects, adapted from [15].

The recorded data revealed variability in participants' use of spatial deixis, both across individuals and within the same arrangements. This variability suggests that, at least in a multi-object selection task, the flexible nature of the peri-personal space [12] could also be influenced by other cognitive or mental representations to the ones mentioned in the introduction.

Our hypothesis is that within a given environment, the choice between *here* and *there* is shaped not only by metric distance, but also by how users perceive the size, extent and organization of the environment. In the present study, the cupboard serves as a controlled spatial setting for examining such contextual influences (cf. [10]). Although the analysis is grounded in this specific setting, we expect the findings to be transferable to similar spaces and scenarios.

3 Conceptual Framing

To account for the observed variability in deixis use, we examine the data from a socio-cognitive perspective. This framing allows us to consider how the use of *here* and *there* is shaped by the way individuals interpret the interaction situation. Since the boundaries of proximal space are dynamically modulated, participants may draw on embodied spatial perspectives [10], prior experiences, and personal beliefs to form internal mental models for the task at hand and to resolve ambiguities [17, 18]. These mental models are inherently unstable and may be revised by the same individual across different instances of similar tasks.

For our analysis, we adopt Norman's distinction [17] between four related components: the target system **(t)**, the conceptual model of that target system **(C(t))**, the user's mental model **(M(t))** and the researcher's conceptualization of the user's mental model **(C(M(t)))** . Within this framing, the present paper interprets the cupboard space as a structured interaction space, and subsequently uses the participants' observed deixis behavior to construct an analytical account of potential user mental models for the use of spatial deixis. Therefore, in terms of Norman's four components, the experimental setting is as follows:

1. **t:** The target system is the shared cupboard space, which a participant and a robot jointly orient towards from fixed positions with a shared perspective.
2. **C(t):** The conceptual model of the target system is the structured representation of the interaction space. In the study, this consists of the bounded cupboard environment and the spatial arrangements of the objects across the five scenarios, as shown in Fig 1. This allows for analyzing how the use of spatial deixis relates to object placement, relative spatial configurations and contextual variation across scenarios.
3. **M(t):** The participants' mental models are assumed to be shaped by embodied reach, object arrangement, attentional selection [6, 12] and contextual interpretation (see next section).
4. **C(M(t)):** The researchers' conceptualization of the participants' mental models is formulated as a hypothesis about how the cupboard space may be interpreted for deixis use. The cupboard is represented as a 3×3 grid of zones. Since participants approach the cupboard from the left, the left column (zones 1, 4, and 7) is hypothesized to correspond more strongly to *here*, the right column (zones 3, 6, and 9) to *there*, and the middle column (zones 2, 5, and 8) to a potentially ambiguous deixis use. In this way, C(M(t)) provides an initial analytical interpretation of how participants may mentally organize the cupboard into proximal, distal and intermediate spatial regions during deixis use.

The next section presents potential mental models, which may underlie the use of a person's spatial deixis in a given context. These models are treated as components of C(M(t)), without any claims of being exhaustive.

4 Methods and results

4.1 Procedure

Our analysis focuses on the verbal spatial deixis produced by the 12 participants for 174 valid trials (out of 180) in [15]. Participants pointed at an object in each trial while verbally referring to it using spatial deixis (e.g., "the cup *there*)." For each participant, this deixis use was examined at the level of individual objects within each scenario, and subsequently compared across scenarios to trace changes in the use of *here* and *there* across zones and

objects. For each object in each scenario, we noted whether it was referred to with *here* or *there*, in which zone of the cupboard it is located (absolute position), its relative closeness to the participant compared to other objects (relative location), and which deixis had been used in references to the object in previous scenarios. We then looked for object-level and scenario-level patterns across participants in order to capture inter-individual variation and to infer possible contextual factors and mental models underlying such variation.

4.2 Analysis

The choice of spatial deixis in a given situation may be shaped by multiple interacting factors, formed consciously or unconsciously. Some of the factors that we consider here are: the **environmental context** (e.g. quantity and arrangement of objects in the cupboard), the **configuration and properties of the objects**, such as relative closeness, and the **human's embodied perspective** relative to the cupboard (e.g. pointing gestures extend participants' peri-personal space). The object size may also influence deixis use, but since the objects in the setup were comparable in size to fit the cupboard, it was excluded from our present analysis.

Considering these factors, we suggest the following potential spatial mental models that enrich **C(M(t))** in the spatial understanding of deixis use:

Absolute model (AM) is characterized as being driven by the contextual environment. In the case of the cupboard setting, the spatial layout itself provides a structured frame that can bias deixis selection as mentioned in C(M(t)) in Section 3.

Relative model (RM) is defined in terms of the relative arrangement of objects within a single scenario. Rather than interpreting an object solely by its absolute proximity to the participant, it examines whether the use of *here* or *there* depends on how that object is positioned relative to other objects in the same arrangement. In this way, RM captures how relational spatial context within a scene may shape deixis use.

Mixed model (MM) is conceptualized as an integration, or interaction, of both absolute and relative mental models.

Temporal model (TM) captures possible carry-over effects in deixis use for an object across scenarios. It examines whether the deixis assigned to the same object in one arrangement is influenced by how that object was interpreted in preceding arrangements. This can be defined as an effect of mental conditioning across scenarios or an adaptability of the deixis use across arrangements.

Response-biased model (RbM) is defined in terms of individual response bias. A participant is considered response-biased if they consistently favor one deixis, for a majority of the objects within a scenario and across scenarios, independent of their absolute and relative positions. Predominant use of *here* might suggest proximal contextual space, within their arm's reach, whereas predominant use of *there* might suggest a more distal space.

4.3 Results

Table 1 shows the applicability of the suggested spatial mental models for each participant (P_id) for all scenarios. Based on these observed patterns, we classify the participants into four groups that characterize their dominant deixis use strategies.

Structured strategy. The first group shows a relatively structured use of spatial deixis across scenarios. These participants primarily seem to follow AM, while also adapting their responses according to the relative positions of nearby objects. This pattern is mostly

Table 1 The table shows the mental models that fit the participants' spatial deixis use in and across scenarios 1 to 5. A tick indicates that the respective mental model may explain deixis use.

P_id.	Mental Models	Scenarios				
		1	2	3	4	5
1	AM	✓	✓	✓	✓	✓
	RM			✓	✓	✓
	MM			✓	✓	✓
	TM		✓	✓	✓	✓
	RbM					
2	AM		✓	✓		
	RM	✓	✓	✓	✓	
	MM		✓	✓		✓
	TM		✓	✓		
	RbM				✓	✓
3	AM	✓	✓	✓	✓	
	RM			✓	✓	
	MM			✓	✓	
	TM		✓	✓	✓	✓
	RbM					✓
4	AM		✓		✓	✓
	RM	✓	✓	✓	✓	✓
	MM		✓		✓	✓
	TM		✓			✓
	RbM					
5	AM	✓	✓			
	RM		✓	✓	✓	
	MM		✓	✓		
	TM		✓		✓	
	RbM				✓	✓
6	AM	✓	✓	✓	✓	
	RM		✓	✓	✓	
	MM		✓	✓	✓	
	TM		✓		✓	
	RbM					
7	AM	✓	✓	✓	✓	✓
	RM		✓	✓	✓	✓
	MM		✓	✓	✓	✓
	TM		✓		✓	✓
	RbM					
8	AM	✓	✓	✓	✓	
	RM		✓	✓	✓	
	MM		✓	✓	✓	
	TM				✓	
	RbM					✓
9	AM	✓	✓	✓	✓	✓
	RM		✓		✓	
	MM		✓		✓	
	TM		✓	✓	✓	✓
	RbM					✓
10	AM	✓	✓	✓		
	RM			✓		
	MM			✓	✓	✓
	TM		✓	✓	✓	✓
	RbM		✓		✓	✓
11	AM	✓	✓	✓	✓	✓
	RM		✓	✓	✓	✓
	MM		✓	✓	✓	✓
	TM				✓	
	RbM					✓
12	AM	✓			✓	
	RM		✓	✓	✓	✓
	MM		✓	✓	✓	
	TM		✓			✓
	RbM			✓		✓

observed for participants 1, 3, 7, 9, and 11. For both participants 1 and 3, the deixis assigned to cup_2 (the cup in zone 2 of the cupboard) remained stable across scenarios, suggesting also a TM effect.

Relative strategy. The second group relies more strongly on RM, where the deixis assigned to an object depends on its relative position to neighboring objects rather than its absolute position on the cupboard. This pattern is clearest for participant 4 but is also observed for participants 7, 8, 11 and 12.

Hybrid strategy. The third group reflects the use of MM, combining both absolute and relative interpretations. This pattern is particularly visible for participants 7 and 11. The responses of participants 2 and 4 suggest sensitivity both to the relative arrangement of objects and to the height of the cupboard. For participant 2, this pattern becomes less stable in scenario 5, whereas participant 4 shows a more structured use even with increased object density.

Variable strategy. The final group shows more inconsistent patterns in their use of spatial deixis across scenarios. For participant 5, only limited local structure seems to be present in scenarios 1 and 2, while participant 10 and 12 shows no clear regularity at all across the scenarios.

In addition, we find the use of RbM in the more complex scenarios 4 and 5. In scenario 5, with increased object density, participants 2, 3, 5, 8, 9, 10, 11, and 12 show a more biased response towards either *here* or *there*, with their use less linked to absolute or relative position of the objects.

5 Discussion

The results suggest that the use of spatial deixis in the cupboard setting cannot be explained by a single stable principle alone. Although the initial hypothesis in C(M(t)) proposes a broad proximal-distal organization of the cupboard as has been found in previous research [4], the analysis of the verbal data indicates that participants do not just rely on this spatial partition in a multi-object arrangement. Instead, their deixis use appears to reflect multiple interacting influences, including absolute position, the relative arrangement of the objects and, in some cases, biases over a specific deixis. The present results differ from previous results in the literature (e.g., [4]), achieved in single-object settings. They support the Norman-inspired framing [17] adopted in this paper, where the participants' latent mental model M(t) may be distinctively different from the experimenters' conceptual model C(M(t)).

From a linguistic perspective, the variability is also consistent with the fact that *here* is not a uniformly simple proximal term. As noted by Fretheim et al. [11], *here* may function as a token-reflexive, anaphoric or deictic expression and its interpretation depends on the contextual information available to the addressee. In this sense, the present analysis is consistent with the view that the use of *here* and *there* is shaped not only by physical metric distances, but also by how the scene is mentally and interactionally organized. The participant groupings further support this interpretation, where participants employed different strategies for their deixis use. Scenario 1 shows the clearest pattern, with 10 out of 12 participants following AM. This suggests that in simpler spatial configurations, people tend to rely on the overall structure of the environment. In contrast, Scenarios 2 to 5 show a combination of different mental models, indicating that increased object density leads to more flexible, context-dependent and less stable deixis use.

The present study offers an initial socio-cognitive interpretation of the observed variability of spatial deixis. Still, there is a need for a more systematic exploration of potential context factors influencing spatial deixis use and data collection from more participants and in other, more systematically selected environment settings in future research. Still, it appears that together, these findings support the idea that the participants' (or speakers') mental model M(t) is flexible and context-sensitive and that the experimenters' (or recipients') mental model C(M(t)) must therefore accommodate more than one strategy for the interpretation of spatial deixis.

This is particularly relevant for designing "artificial" recipients, such as social robots, and for human-robot interaction, where predefined spatial zones may fail to capture the variability in human spatial language, especially when object arrangements are dense or ambiguous. Overall, our study supports a view of spatial deixis being shaped as a combination of contextual and individual factors, mentally organized and variable across individuals.

References

1 Hissah Nasser Alothman and Haroon N Alsager. Acquisition of l2 english spatial deixes by arabic-speaking children. *Frontiers in Psychology*, 13:997110, 2022.

2 Rory J. Bufacchi and Gian Domenico Iannetti. An action field theory of peripersonal space. *Trends in Cognitive Sciences*, 22(12):1076–1090, 2018. doi:10.1016/j.tics.2018.09.004.

3 Karl Bühler. *Sprachtheorie*, volume 2. Jena Fischer, 1934.

4 Michela Caldano and Kenny R. Coventry. Spatial demonstratives and perceptual space: To reach or not to reach? *Cognition*, 191:103989, 2019. doi:10.1016/j.cognition.2019.06.001.

5 Malinda Carpenter and Michael Tomasello. Joint attention and imitative learning in children, chimpanzees, and enculturated chimpanzees. *Social Development*, 4(3):217–237, 1995.

6 Marvin M Chun, Julie D Golomb, and Nicholas B Turk-Browne. A taxonomy of external and internal attention. *Annual review of psychology*, 62(1):73–101, 2011.

7 Holger Diessel and Kenny R Coventry. Demonstratives in spatial language and social interaction: An interdisciplinary review. *Frontiers in Psychology*, 11:555265, 2020.

8 Nick J Enfield. Demonstratives in space and interaction: Data from lao speakers and implications for semantic analysis. *Language*, 79(1):82–117, 2003.

9 Charles J Fillmore. Santa cruz lectures on deixis, 1971. *(No Title)*, 1975.

10 Christian Freksa, Alexander Klippel, and Stephan Winter. A Cognitive Perspective on Spatial Context. In Anthony G. Cohn, Christian Freksa, and Bernhard Nebel, editors, *Spatial Cognition: Specialization and Integration*, volume 5491 of *Dagstuhl Seminar Proceedings (DagSemProc)*, pages 1–16, Dagstuhl, Germany, 2007. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/DagSemProc.05491.4.

11 Thorstein Fretheim, Nana Aba Appiah Amfo, and Ildikó Vaskó. Token-reflexive, anaphoric and deictic functions of "here". *Nordic Journal of Linguistics*, 34(3):239–294, 2011. doi:10.1017/S0332586512000030.

12 Harmen B Gudde, Jacqueline Collier, and Kenny R Coventry. Spatial demonstratives and perspective taking in english and japanese. *Cognitive Science*, 50(3):e70183, 2026.

13 Edward T Hall. Proxemics—the study of man's spatial relations. *Man's image in medicine and anthropology*, 4:1003–1026, 1963.

14 William F Hanks. *Referential practice: Language and lived space among the Maya*. University of Chicago Press, 1990. doi:10.1017/S0047404500017036.

15 Adwitiya Mandal, Kai-Florian Richter, and Zoe Falomir. How robots understand 'here' and 'there': A perceptual model for spatial deixis. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, HRI Companion '26, pages 1018–1022, New York, NY, USA, 2026. Association for Computing Machinery. doi:10.1145/3776734.3794548.

16 Sandra Murphy. Children's comprehension of deictic categories in oral and written language. *Reading Research Quarterly*, pages 118–131, 1986.

17 Donald A Norman. Some observations on mental models. In *Mental models*, pages 7–14. Psychology Press, 2014.

18 Jane Oakhill and Alan Garnham, editors. *Mental models in cognitive science: essays in honour of Phil Johnson-Laird*. Psychology P., 1996.

19 Amalia Skilton. The deictic content of demonstratives. *Language and Linguistics Compass*, 18(4):e12519, 2024. doi:10.1111/lnc3.12519.

Document Geocoding for Inferring Spatial Familiarity Within Reddit Text

Paddy Smith
School of Geography, University of Leeds, UK

Myles Gould
School of Geography, University of Leeds, UK

Ed Manley
School of Geography, University of Leeds, UK

Abstract

In this paper, we reframe social media user geolocation as a task of inferring a user's spatial familiarity. We propose an LLM-based approach that identifies places a user is most familiar, including the strength of familiarity. We evaluate this approach using two Reddit evaluation corpora constructed from subreddit user-flair tags: Reddit-London, where users are associated with London boroughs, and Reddit-USA, where users are associated with US states. We compare these datasets with GeoText benchmark dataset. The results show that LLMs can recover place-associated labels from social media text, achieving the strongest performance on Reddit-USA state prediction and moderate performance on fine-grained London borough prediction. We further show that inferred familiar places exhibit clear spatial dependency. Using polygon-based metrics, familiar places inferred from Reddit-London users have a mean distance to labeled boundary of 3.48 km, while Reddit-USA shows stronger alignment with state labels compared to GeoText, with higher inside-label share, and lower spatial dispersion. These findings suggest spatial familiarity provides a useful alternative framing for social media user geolocation, and that Reddit flair-based labels are more closely aligned with expressed spatial familiarity than home location labels in GeoText benchmark dataset.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases spatial familiarity, document geocoding, geoparsing, LLMs

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.25

Category Short Paper

Declaration on GenAI/LLM use During the preparation of this work, the authors used GPT-5.5, for improving structure of manuscript and supporting the coding of experiments. Use of Gen-AI and has been reviewed and verified by the authors.

1 Introduction

Identifying the location of social media users is useful across a range of domains, including disaster response, public health [11] and event detection [15]. In urban research, such approaches have been used to examine mobility patterns [9], social ties [17] and mapping place boundaries [18]. With the sparsity of available geotagged social media data [4], and the unreliability of self-declared location [8], there is a large body of literature that attempts to infer user location automatically from text [10, 6, 13].

A common approach to identifying social media user location is document geocoding, in which user-generated text is combined into a single document, and then analysed to infer a geographic location, typically interpreted as "home location" [12]. The primary assumption is that users will employ location or geo- indicative words, such as dialects, place names, cultural references, that implicitly or explicitly encode an association with a particular location [6, 1].

© Paddy Smith, Myles Gould, and Ed Manley;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 25; pp. 25:1–25:10

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This approach assumes that the inferred location corresponds to where the user resides, based on their social media activity. However, users may not explicitly mention location-indicative terms in posts, or refer to places from locations they do not reside in [16]. Place references may be made for a variety of reasons, such as travel, past experiences or personal interests. For example, frequent references that indicate Boston, USA, may reflect support for a sports team, rather than physical location in the city. Despite this, many ground-truth datasets are created based on this assumption, such as GeoText [3] and Twitter-US [14], where user location is defined by the geotag of their first tweet [16]. The reliance on "home location" can therefore produce misleading conclusions about the relationship between text and user location, impacting models trained and evaluated on such data.

This limitation is more pronounced in discussion-oriented platforms such as Reddit, where users can engage in discussions, often without explicitly mentioning their location. Unlike other social media platforms, Reddit lacks geotagged data and often contains rich, descriptive conversations about place that reflect varying levels of familiarity. In this context, inferring a single "home location" is often neither feasible or appropriate. Instead, it is more meaningful to consider a user's spatial familiarity, meaning a *"close acquaintance with an environment or its elements"* [5]. It is about the extent a place is known [2], through direct exposure via regular interactions, or indirect exposure to information about places [5].

In this paper, we reconceptualise social media user geolocation as a task of inferring user's familiarity with place, rather than predicting a single home location. We propose an LLM-based approach that identifies a users familiar places across multiple spatial scales, along with the strength of this familiarity. We examine the spatial distributions of users spatial familiarity, evaluating the extent to which they exhibit spatial dependency. We evaluate this approach using proxy ground-truth signals derived from user-generated metadata (i.e., subreddit user flair) and a benchmark GeoText dataset.

The main contributions of this paper are threefold: (1) we reconceptualise social media user geolocation as spatial familiarity inference rather than home location; (2) we demonstrate a methodology for constructing spatial familiarity evaluation corpora from subreddit flair tags; and (3) we provide evidence that spatial familiarity extracted from social media users exhibits spatial dependency, with Reddit text showing tighter spatial clustering around ground-truth labels than GeoText home location labels.

2 Methodology

2.1 Reddit Data

Evaluating LLMs ability to infer Reddit user's spatial familiarity is challenging, due to the lack of ground-truth data [7]. Currently, there are no datasets that have gold-standard location for Reddit comments, or users. In addition, there is a very limited amount of location data attached to Reddit, with users not required to input their own location.

In order to address this, we utilise the "user flair" tag, which are self-attributed labels that are specific to individual subreddits. Often in place-based subreddits, these flairs indicate a user's association with a particular place. For example, in r/London it is common for users to tag a London borough, whilst r/AskAnAmerican users may tag their associated US state. Although these tags do not guarantee the user is located within this area, it can be presumed that the user has a high likelihood of familiarity with this place.

To build a spatial familiarity evaluation corpora, we collect Reddit comments and posts from r/London and r/AskAnAmerican subreddits using the Reddit API[1]. These subreddits are chosen as they provide opportunity to evaluate spatial familiarity inference at two different spatial scales. Users are identified if they have a London borough or US state as their flair tag, or a place within a London borough or state which is subsequently matched manually. Then, comments and posts are collected across the subreddit, and other subreddits they may have contributed to (i.e., r/LondonCycling, r/NewEngland). For each user, their comments and posts are combined together to form a document. Reddit user documents are cleaned to remove unusual characters and whitespaces in the text, and truncated to 35,000 characters, to fit within the LLM context window.

This produces two Reddit-datasets; Reddit-London containing 204 users with London borough labels, and Reddit-USA, containing 239 users with US state labels. LLMs are then evaluated based on their ability to identify the London borough or US state in which the user is associated in (via their flair tag) using the user document. As this is a classification task, precision, recall and F1-score are used to evaluate model performance.

2.2 GeoText

GeoText, a widely used benchmark data within social media user geolocation tasks [10], is also used in this study. We evaluate the LLM-based approach using Geotext, due to the sparsity of benchmark datasets available. GeoText is compromised of 9,475 Twitter users from the US [3]. The location of the user is defined as the geographic coordinates of their first Tweet. GeoText has typically been used to evaluate social media user geolocation methods, by computing distance error metrics between the gold-standard coordinates and predicted coordinates. However, distance-based metrics do not capture if LLMs are able to identify levels of user spatial familiarity. Therefore, we adapt GeoText by assigning each user to a US state, by intersecting their location with state boundaries. An LLM must then predict a US state that the user is based within. The assumption is that the user would have some familiarity with the state, and the LLM can infer these based on the geo-indicative words and place mentions.

2.3 Model inference

We implement two open-source LLMs; `qwen2.5-72B`[2] and `Llama3.3-70B-Instruct`[3], that have been quantized and are available via HuggingFace. We also deployed `claude-sonnet-4.6`[4] on the Reddit-based datasets to test a larger parameter model. Two different prompt are constructed, giving instructions to estimate spatial familiarity for Reddit-London and Reddit-USA (Appendix A.1) and home location for GeoText (Appendix A.2). For this paper, only zero-shot prompt approaches are experimented with. The model is instructed to output in a predefined JSON schema. Each user is assigned a predicted location, a familiarity score, and then a list of places they are familiar with.

[1] https://redditinc.com/policies/data-api-terms
[2] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-AWQ
[3] https://huggingface.co/casperhansen/llama-3.3-70b-instruct-awq
[4] https://www.anthropic.com/news/claude-sonnet-4-6

2.4 Evaluating spatial dependency

During model inference, the models also predict a set of familiar places for each user, including the location of these places and an estimated familiarity score. These inferred places are used to assess whether a user's place familiarity is spatially dependent on their ground-truth location labels. The spatial dependency analysis uses polygon-based distance metrics. This is appropriate because the ground-truth labels in this study refer to administrative areas, namely London boroughs and US states. For each inferred familiar place, we calculate its distance to the user's labeled administrative boundary. Places located inside the labeled borough or state are assigned a distance of zero, while places outside the labeled area are assigned their distance to the nearest point on the boundary.

For each user, we calculate polygon mean distance to anchor (polygon MDA), defined as the mean distance between a user's inferred familiar places and their labeled polygon. We also calculate a share of familiar places located inside the labeled area and the mean distance of outside-label places from the labeled boundary. Lower polygon distance and higher inside-label share indicate stronger spatial dependency between inferred place familiarity and user's ground-truth label. Additionally, spatial dispersion is computed as the mean pairwise distance between all familiar places for a given user. Higher dispersion indicates a user's familiar places are spread across a wider geographic area, whereas lower dispersion indicates concentrated local familiarity.

Finally, kernel density estimation (KDE) is used to visualize user's spatial familiarity on different aggregate spatial scales. For each user, the LLM-extracted familiar places are aggregated to produce a density surface, weighted by the familiarity score assigned to each place.

3 Results

3.1 Predicting user location labels

We first predict user's location labels for both Reddit-London, Reddit-USA and GeoText datasets. For Reddit-London and Reddit-USA, we are predicting the associated London borough or US state, whereas for GeoText we are predicting the US state as a home location. Model performance for this task is shown in Table 1.

Table 1 Model evaluation for predicting ground-truth location labels for Reddit-London, Reddit-USA and GeoText users.

Model	Dataset	Resolved Users	Precision	Recall	F1-score
Qwen2.5-72B	Reddit-London	99%	0.63	0.5	0.52
Llama3.3-70B	Reddit-London	94%	0.57	0.48	0.5
Claude-Sonnet-4.6	Reddit-London	90%	0.71	0.58	0.61
Qwen2.5-72B	Reddit-USA	93%	0.78	0.78	0.77
Llama3.3-70B	Reddit-USA	96%	0.86	0.88	0.85
Claude-Sonnet-4.6	Reddit-USA	100%	0.9	0.89	0.88
Qwen2.5-72B	GeoText	76%	0.42	0.33	0.36
Llama3.3-70B	GeoText	66%	0.46	0.37	0.39

High F1-scores are achieved across the models for the Reddit-USA dataset. Therefore the task of predicting the US state flair-tag based on the Reddit user document was performed accurately and with minimal errors. In contrast, the same models performed with less precision

when predicting US states for the GeoText users. This may highlight the misalignment between the user document and the assigned home location in the GeoText dataset. For the Reddit-London dataset, the models achieved modest precision and recall scores. Despite the granularity of London boroughs, the models were relatively successfully in correcting labeling a London borough for the Reddit user. As expected, the larger parameter `claude-sonnet-4.6` shows the highest performance on the Reddit-London and Reddit-USA datasets.

3.2 Spatial dependency

Next, we examine whether the familiar places inferred for each user are spatially dependent on the user's ground-truth location label using polygon-based metrics (Table 2). Across all datasets, Reddit-London showed the lowest polygon MDA (3.48km) and outside-boundary distance (4.47km), however had a lower inside-label share (33%). Therefore often familiar places mentioned by a user are nearby to their assigned London borough, but not necessarily within.

Table 2 Polygon-based spatial dependency metrics for Reddit-London, Reddit-USA and GeoText.

Dataset	Polygon MDA (km)	Inside-label share (%)	Outside-boundary distance (km)	Spatial dispersion (km)
Reddit-London	3.48	33	4.47	5.49
Reddit-USA	164.28	80	291.41	263.03
GeoText-state	890.12	49	1081.12	695.34

Reddit-USA shows a different pattern, with a much larger share of inferred familiar places inside labeled states (80%). However, polygon MDA and outside-boundary distance are substantially larger than for Reddit-London, reflecting the broader geographic scale of US states. This indicates stronger within-label alignment, but a much larger spatial extent for familiar places.

GeoText can be compared more directly with Reddit-USA, as they both have US states as user location labels. GeoText shows a weaker spatial dependency, with a lower inside-label share (49%), higher polygon MDA (890.12 km), and larger outside-boundary distance (1081.12 km). GeoText also has a higher spatial dispersion, suggesting familiar places inferred from GeoText users are more geographically spread out. These results support the argument that Reddit flair-based labels are more closely aligned with expressed place familiarity, while GeoText home-location labels capture a different and less spatially concentrated relationship between users and places.

We next use KDE visualizations to examine the spatial structure underlying the spatial dependency analysis. For Reddit-London, we show boroughs that are most and least spatially clustered in terms of familiar places (Figure 1). For Reddit users associated with Islington and Ealing, familiar places commonly discussed are more spatially concentrated within or around the borough boundary. In comparison, users associated with Waltham Forest and Bromley commonly reference more dispersed familiar places, spanning multiple boroughs and extending into central London. This helps explain why Reddit-London has a low polygon MDA despite a relatively low inside-label share, as many familiar places fall outside the labeled borough, but remain nearby within the wider urban context.

We then compare KDEs for selected US states across Reddit-USA and GeoText (Figure 2). States are selected where both datasets contain sufficient users and familiar place mentions. Comparing the datasets, Reddit-USA familiar places are substantially more concentrated within or near the labeled state boundary. In contrast, GeoText familiar places are more

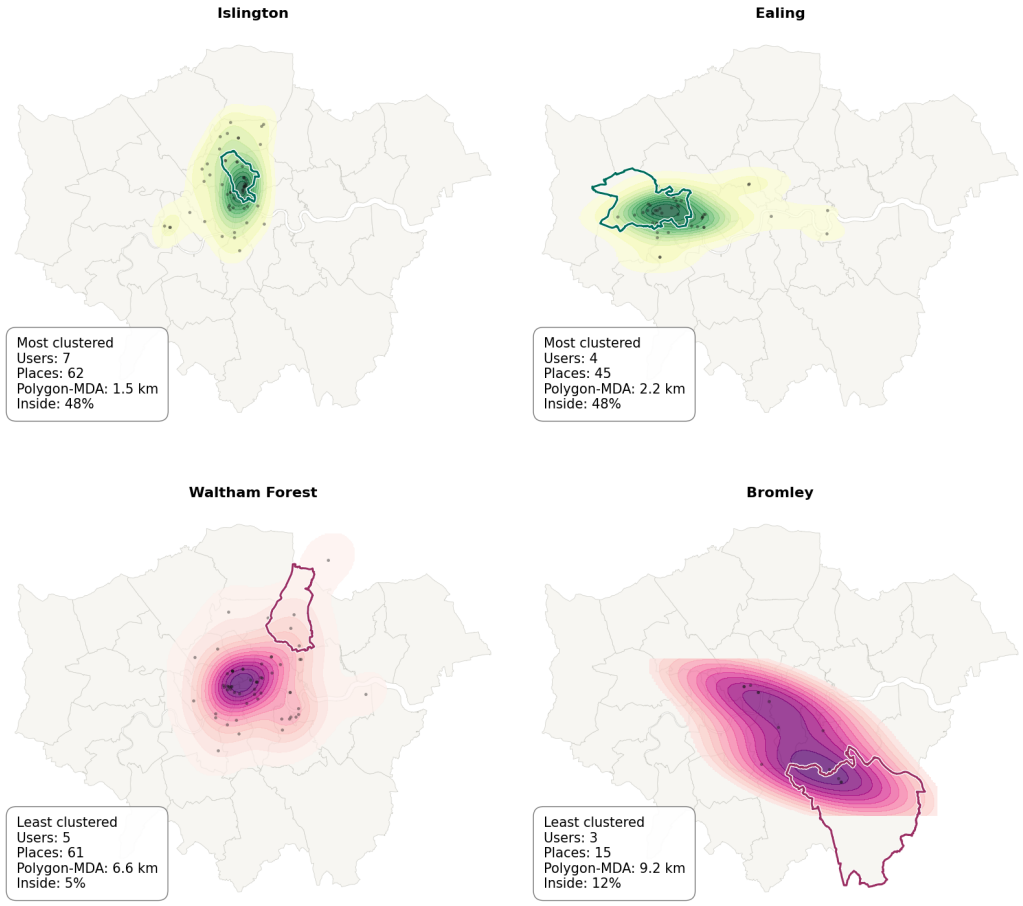


Figure 1 Spatial familiarity density for selected London boroughs. KDE surfaces are weighted by LLM familiarity scores, with each point denoting an inferred familiar place.

spatially dispersed and less consistently aligned with the assigned home-state label. For example, users assigned to Texas in GeoText reference familiar places across a wider national geography, including California and New York. Whereas, Reddit-USA users show a stronger connection of familiar places within Texas. These patterns support the polygon-based metrics, which show higher inside-label share and lower outside-boundary distance for Reddit-USA compared to GeoText.

4 Discussion & Conclusion

In this short paper we have considered the limitations of identifying "home location" of social media users, and instead demonstrated an alternative spatial familiarity approach. The potential application of this approach is demonstrated on Reddit data, a social media dataset which lacks user location information, however is increasingly being used for urban analysis. We illustrate an approach to constructing a place familiarity evaluation corpora using subreddit user flair-tags. We provide preliminary evidence that LLMs can support fine-grained user location inference (i.e., London boroughs), as opposed to coordinates or larger spatial units. We also demonstrate the potential for LLMs estimating continuous levels of spatial familiarity and identifying sets of locally relevant places associated with individual users.

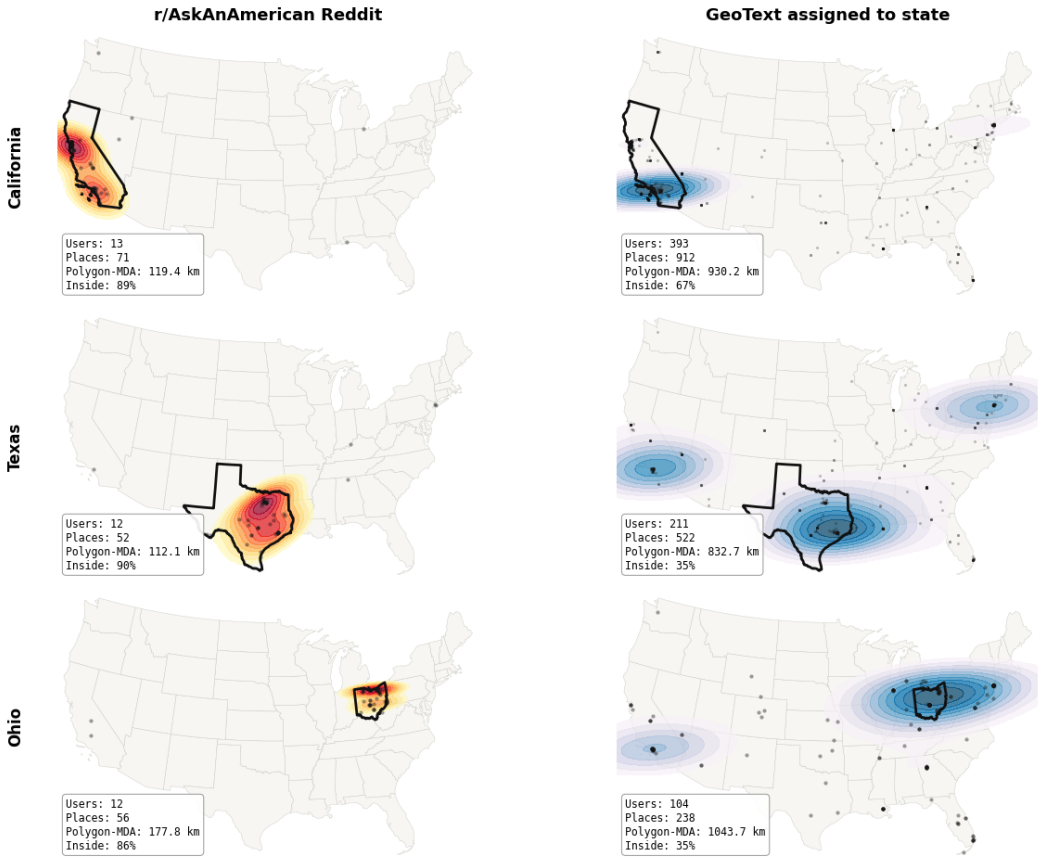


Figure 2 Spatial familiarity density for selected US states in Reddit-USA and GeoText. KDE surfaces are weighted by LLM familiarity scores, with each point denoting an inferred familiar place.

Further work is required to validate model inferences to assess how closely LLM predictions align with actual spatial familiarity. In particular, the Reddit-based evaluation corpus presents several limitations. Firstly, the flair-tags are self-reported therefore unverified and might not reflect an individuals own familiarity. Secondly, the size of our evaluation corpora is relatively small, providing opportunity to increase dataset size by incorporating other place-based Subreddits (r/LosAngeles, r/Australia). Finally, the familiarity scores and locally relevant places are not independently verified, therefore cannot assess if they are meaningful. Future work could compare these scores to survey-based scores, or other urban perception datasets.

We also demonstrate that Reddit user spatial familiarity exhibits clear spatial dependency, forming coherent geographic patterns. Reddit-USA shows a polygon mean distance to anchor of 164.28km, compared to 890.12km for GeoText users, suggesting that social media text from Reddit closer reflects the geographic context indicated by the ground-truth label. In contrast, the higher spatial dispersion in GeoText challenges the validity of home location as a ground-truth proxy, as the geographic content of user text frequently does not align with the registered home location. These findings suggest that Reddit flair-based labels may represent a higher quality ground-truth signal for place familiarity evaluation than geotag-derived home locations.

References

1. Benjamin Adams and Krzysztof Janowicz. On the Geo-Indicativeness of Non-Georeferenced Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):375–378, August 2012. doi:10.1609/icwsm.v6i1.14309.
2. David J. Chalmers and Robert G. Knight. The Reliability of Ratings of the Familiarity of Environmental Stimuli: A Generalizability Analysis. *Environment and Behavior*, 17(2):223–238, March 1985. doi:10.1177/0013916585172003.
3. Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.
4. Deen Freelon. Computational Research in the Post-API Age. *Political Communication*, 35(4):665–668, October 2018. doi:10.1080/10584609.2018.1477506.
5. Nathan Gale, Reginald G. Golledge, William C. Halperin, and Helen Couclelis. EXPLORING SPATIAL FAMILIARITY. *The Professional Geographer*, 42(3):299–313, August 1990. doi:10.1111/j.0033-0124.1990.00299.x.
6. Bo Han, Paul Cook, and Timothy Baldwin. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. In *Proceedings of COLING 2012: Technical Papers, Pages 1045–1062,*, 2012.
7. Keith Harrigian. Geocoding Without Geotags: A Text-based Approach for reddit. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 17–27, Brussels, Belgium, 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-6103.
8. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246, Vancouver BC Canada, May 2011. ACM. doi:10.1145/1978942.1978976.
9. Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding Human Mobility from Twitter. *PLOS ONE*, 10(7):e0131469, July 2015. doi:10.1371/journal.pone.0131469.
10. Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. An overview of microblog user geolocation methods. *Information Processing & Management*, 57(6):102375, November 2020. doi:10.1016/j.ipm.2020.102375.
11. Hanjia Lyu, Long Chen, Yu Wang, and Jiebo Luo. Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19. *IEEE Transactions on Big Data*, 2020. doi:10.48550/arXiv.2004.06307.
12. Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–21, October 2014. doi:10.1145/2528548.
13. Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting Text and Network Context for Geolocation of Social Media Users, 2015. doi:10.48550/arXiv.1506.04803.
14. Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.
15. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, Raleigh North Carolina USA, April 2010. ACM. doi:10.1145/1772690.1772777.
16. Zhaomin Xiao and Eduardo Blanco. Are People Located in the Places They Mention in Their Tweets? A Multimodal Approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2022.

17 Sarita Yardi and Danah Boyd. Tweeting from the Town Square: Measuring Geographic Local Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):194–201, May 2010. doi:10.1609/icwsm.v4i1.14016.

18 Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, July 2017. doi:10.1080/13658816.2017.1282615.

A LLM Prompts

A.1 Prompt for estimating place familiarity within Reddit-London

```
You are analysing Reddit comments written by a single user from London related subreddits.
Your goal is to estimate levels of place familiarity, which is the extent an
individual knows their surroundings. Familiarity with a place does not necessarily mean a
user resides in that place.
### TASK
Assess the users level of familiarity with London, and identify specific London borough.
Additionally, identify any specific local places (e.g., stations, parks, landmarks) the
user appears familiar with. For each such place, provide its name, a familiarity score
(0-1), and its approximate coordinates (latitude, longitude).
There is no limit on the number of places you may list.
### How to identify familiarity?
- Repeated references to specific neighbourhoods or places within a borough
- Detailed local knowledge (routes, transport, pricing, safety, comparisons)
- Personal or experiential language ("I live near...", "my area...", "I always go to...")

Return ONLY valid JSON in this format:
{{
  "PredictedBorough": "BOROUGH_NAME",
  "ApproximateLocation": {{
    "lat": 0.0,
    "lon": 0.0
  }},
  "CityFamiliarityScore": 0.0,
  "LocalPlaceFamiliarity": [
    {{
      "place": "PLACE_NAME",
      "familiarity": 0.0,
      "lat": 0.0,
      "lon": 0.0
    }}
  ],
  "Confidence": 0.0,
  "Evidence": "brief explanation of decision"
}}
User comments:
{document}
```

A.2 Prompt for estimating home location of Twitter users in GeoText

```
You are analysing Tweets written by a single user in the United States.
TASK: Predict which state the user most likely lives in.
```

```
Return ONLY valid JSON in this format:
{{
    "PredictedState": "STATE_NAME",
    "Confidence": 0.0
}}
Rules:
- Confidence must be between 0 and 1
- Output JSON only
- No explanation or markdown

User comments:
{document}
```

Learning a Map Under Positive Emotions: How Self-Reported Arousal and Individual Differences Relate to Route and Survey Knowledge

Veronica Muffato[1]
Department of General Psychology, University of Padova, Italy

Laura Miola
Department of General Psychology, University of Padova, Italy

Francesca Pazzaglia
Department of General Psychology, University of Padova, Italy

Chiara Meneghetti
Department of General Psychology, University of Padova, Italy

Abstract

Digital map learning is a common everyday activity requiring the encoding of landmarks and their spatial relations, and may be related to individual differences. The present study investigated how individual positive emotional states differing in arousal (high vs. low) relate to route and survey knowledge after spatial learning from maps. A total of 156 participants (ongoing data collection) completed an online within-subject procedure including emotionally induced scenarios, map study, and recall tasks assessing route (landmark order) and survey (landmark location task) knowledge. Results showed no direct effect of arousal condition on performance; however, individual increases in arousal predicted lower route knowledge accuracy, while survey knowledge was unaffected. Importantly, trait positive affect positively predicted both route and survey performance. Age negatively predicted both, gender differences emerged only for survey knowledge, whereas spatial anxiety impaired route knowledge. These findings suggest a selective role of arousal in route processing and highlight the importance of considering individual differences in spatial learning from maps.

2012 ACM Subject Classification Social and professional topics → User characteristics

Keywords and phrases map, emotions, arousal, route knowledge, survey knowledge, spatial cognition

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.26

Category Short Paper

1 Introduction

Reading maps to acquire environmental information is a common experience that requires processing and memorizing details about landmarks and their spatial relations. Spatial learning from maps is related to a variety of individual and contextual factors. One under-explored individual factor is the learner's emotional state, which may affect the ability to form accurate spatial mental representations. Consider, for example, reading a map to find the route to a hospital while feeling rushed and stressed, versus reading a map as a relaxed tourist exploring an unfamiliar city. Both situations involve spatial learning from a map, yet they occur under markedly different emotional conditions.

[1] Corresponding author

© Veronica Muffato, Laura Miola, Francesca Pazzaglia, and Chiara Meneghetti;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 26; pp. 26:1–26:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Emotions can be conceptualized along two primary dimensions: valence (i.e., the positivity or negativity of an emotion) and arousal (i.e., the intensity of emotional activation), ranging from negative–low arousal states (e.g., sadness) to negative–high arousal states (e.g., anxiety), and from positive–low arousal states (e.g., calmness) to positive–high arousal states (e.g., excitement). Previous research has shown that emotional states, defined along these dimensions [5, 20], are associated with memory and cognitive processing [1, 10, 21]. A growing body of research suggests that emotional factors are closely linked also to spatial memory and the formation of spatial mental representations, particularly following navigation in real or virtual environments. Overall, negative emotional states, especially those characterized by high arousal such as stress, tend to impair spatial processing. For instance, time pressure reduces the accuracy of cognitive maps (e.g., [8, 24, 23, 25]), while stress increases reliance on familiar routes rather than novel shortcuts, leading to longer travel distances and increased task completion times (e.g., [4]). More broadly, negative valence combined with high arousal has been associated with reduced spatial flexibility and distorted spatial judgments (e.g., [3, 6]). In contrast, positive emotional states, particularly at low levels of arousal, are generally associated with improved spatial recall (e.g., [16, 18, 19, 17]. Additionally, arousal itself, independently of valence, can relate to spatial representations by affecting perceived distances and the relationships between landmarks (e.g., [7, 11]). However, to our knowledge, no study to date has specifically investigated the role of emotions in spatial learning from a map and recall of both route and survey knowledge. Route knowledge refers to a sequential, egocentric representation of the environment, based on the ordered series of actions and landmarks encountered along a path. In contrast, survey knowledge reflects an allocentric, map-like representation that encodes the spatial relationships between locations. Understanding how emotional states relate to these distinct forms of spatial knowledge may provide important insights into the mechanisms underlying spatial learning from a map and this represents the research gap addressed by the present study.

Therefore, the present study aims to investigate the role of emotions in the recall of route and survey information following spatial learning from a map under different emotional conditions. Specifically, we focus on the effect of arousal levels (high vs. low) on spatial recall while keeping valence positive, given that positive emotions appear to be the most favourable condition for spatial memory.

In addition, individual differences, such as age, gender, trait levels of positive and negative affect, and spatial anxiety, were considered as potential factors related both to the recall of spatial information learned from a map and to emotional experience. Importantly, the study adopted a within-participant design.

2 Method

Participants and procedure

The study involved 156 participants (103 women and 53 men; data collection is ongoing), aged between 18 and 70 years (M = 37.13, SD = 16.03). Participants were recruited through word of mouth by the experimenters.

They completed the questionnaires and tasks online via Qualtrics in approximately 25 minutes. After reading and providing informed consent, participants reported their emotional valence and arousal using the digital Affective Slider (AS; [2]) before reading and imagining themselves in a described scenario. They were subsequently presented with a positively valenced situation. Two situations were created, and their presentation was randomized across participants. Each situation was designed to elicit a different level of arousal. In

the high-arousal condition (e.g., excitement, enjoyment), the scenario involved being in an unfamiliar city in which participants aimed to memorize a route from a map in order to reach a restaurant to meet old friends. In the low-arousal condition (e.g., calmness, relaxation), the scenario involved being in an unfamiliar city in which participants aimed to memorize a route from a map in order to reach a park to relax. After reading the scenario, participants again reported their emotional valence and arousal using the Affective Slider (therefore, post–pre difference scores were calculated for both self-reported valence and arousal). Participants were then presented with a fictitious city map. Two ad hoc schematic urban maps (960 × 720 pixels) were used as experimental stimuli. Each map depicted a comparable network of city streets and a predefined route marked by a red dashed line. The maps were designed using a simplified cartographic style inspired by common web-based mapping services. Each map contained 14 landmarks distributed along the route. Landmarks were represented by pinpoint symbols and identified through verbal labels. The selection and placement of landmarks were validated in a pilot study to ensure comparable levels of recognizability across maps. Following a filler task, two spatial information recall tasks were administered in random order: landmark order task (route knowledge) and landmark location task (survey knowledge). In the Landmark order task participants were asked to recall the landmarks in the correct order along the path (one point was awarded for each landmark recalled in its correct position). In the Landmark location task participants were asked to identify the correct positions of landmarks on the map by choosing among distractor options (one point was awarded for each landmark location correctly identified). Participants then completed another filler task before rating their valence and arousal again, both before and after reading the second description of situation. Then, after a filler task, participants learned the second city map and completed the Landmark order task and Landmark location task. Then, as control participants were asked to recall the two situations presented and their self-reported level of immersion in each, defined as how much they were able to empathize with the described situation. Finally, participants reported their trait positive and negative affect (PANAS; [22]), spatial anxiety (Spatial Anxiety Scale, SAS; [9], adapted from [12]) and their demographic information (e.g., age, gender).

3 Results (collected sample)

First, we verified that the scenarios effectively increased positive emotional states by comparing post- and pre-scenario ratings. For valence, increases in positive valence were observed in both conditions (high-arousal condition: M = 9.68, SD = 19.15; low-arousal condition: M = 5.01, SD = 18.56), with a greater increase in the high-arousal condition (p = .030, d = 0.25). Arousal also increased in the high-arousal condition (M = 7.41, SD = 16.47) and in the low-arousal condition (M = 2.13, SD = 19.61), with a greater increase in the high-arousal condition (p = .010, d = 0.29), indicating that the manipulation of arousal was effective. Furthermore, the two conditions did not differ in reported immersion (high-arousal: M = 2.78, SD = 1.03; low-arousal: M = 2.65, SD = 1.01; p = .268).

Then, we ran linear models on the currently collected sample, using accuracy on the Landmark Order Task (route knowledge) and the Landmark location task (survey knowledge) as dependent variables. Predictors included type of situation (high vs. low arousal), change in state valence (post–pre), change in state arousal (post–pre), gender, age, trait positive affect, trait negative affect, spatial anxiety, and perceived immersion.

Concerning the Landmark Order Task (route knowledge), no effect of high vs. low arousal condition was observed. However, change in arousal (post–pre) emerged as a significant negative predictor ($\beta = -.13$, $p = .020$), indicating that greater increases in arousal were

associated with lower route knowledge task performance. Trait positive affect (PANAS-PA) was positively associated with performance ($\beta = .19$, $p = .001$), suggesting that higher levels of positive trait affect predicted better route knowledge. In addition, performance decreased with age ($\beta = -.13$, $p = .026$). No interaction emerged between emotional condition and age, indicating that this relationship did not vary as a function of age. No gender differences were observed. Finally, spatial anxiety was negatively associated with route knowledge ($\beta = -.13$, $p = .027$), indicating that higher spatial anxiety was linked to poorer route knowledge after spatial learning from a map.

Concerning the Landmark location task (survey knowledge), no effect of high vs. low arousal condition was observed, nor was any effect of change in arousal (post–pre). Trait positive affect (PANAS-PA) was positively associated with performance ($\beta = .13$, $p = .028$), indicating that higher levels of positive affect predicted better survey knowledge. Age was negatively associated with performance ($\beta = -.12$, $p = .035$). Gender also emerged as a significant predictor ($\beta = -.50$, $p < .001$), indicating that women showed lower performance than men. However, no interaction emerged between emotional condition and both age and gender.

4 Discussion and Conclusion

Overall, the results suggest that emotional factors play a selective role in spatial learning from maps. The manipulation of arousal did not directly affect performance in either task, but individual increases in arousal (post–pre) were associated with poorer route knowledge, in line with evidence that higher arousal can interfere with sequential spatial processing (e.g., [3]). No such effect emerged for survey knowledge. This pattern may be due to the nature of spatial learning from maps, in which survey information is directly supported by the visual format of the map, providing an explicit representation of spatial relationships. As a result, survey knowledge may be less sensitive to fluctuations in arousal and more strongly related to other factors (e.g., visuospatial working memory, [13]). Importantly, trait positive affect was positively associated with performance in both tasks, consistent with theories linking positive affect to enhanced memory [10], and extending this evidence to spatial memory from a map. The negative association between age and performance across tasks is consistent with previous evidence of age-related decline in spatial knowledge after learning spatial information from a map (e.g., [14, 26]). Similarly, gender differences emerged only for survey knowledge, with women showing lower accuracy, in line with prior findings (e.g., [15]), and extending this pattern to spatial recall from maps. Therefore, age and gender have been confirmed as factors to consider when investigating environmental learning. Finally, spatial anxiety was negatively associated with route knowledge, suggesting a specific impact of spatial anxiety on sequential spatial knowledge [12]. Although the present study provides valuable insights, its main limitation is the lack of psychophysiological measures, which could not be collected in the context of an online study. Future research should address this limitation by incorporating objective physiological indices of emotional responses. We also note that the present study is part of a broader research project. An additional study is currently investigating the role of arousal in negatively valenced emotional states (e.g., sadness vs. anxiety) to provide a more comprehensive understanding of the relationship between emotion and spatial learning from maps.

To conclude, emotions elicited in users while they are learning a map may impact their subsequent recall of spatial (route) information. A deeper investigation into the intriguing association between emotion and spatial knowledge after learning a map is therefore important, in order to understand the impact of user-specific factors.

References

1. F. Gregory Ashby. A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, 106(3):529–550, 1999. doi:10.1037/0033-295x.106.3.529.
2. Alberto Betella and Paul F. M. J. Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLOS ONE*, 11(2):e0148037, February 2016. doi:10.1371/journal.pone.0148037.
3. Christophe Blaison and Ursula Hess. Affective judgment in spatial context: How places derive affective meaning from the surroundings. *Journal of Environmental Psychology*, 47:53–65, September 2016. doi:10.1016/j.jenvp.2016.05.005.
4. Alexander P. Boone, Tom Bullock, Mary H. MacLean, Tyler Santander, Jamie Raymer, Alexander Stuber, Liann Jimmons, Gold N. Okafor, Scott T. Grafton, Michael B. Miller, Barry Giesbrecht, and Mary Hegarty. Resilience of navigation strategy and efficiency to the impact of acute stress. *Spatial Cognition & Computation*, 24(3):195–226, July 2024. doi:10.1080/13875868.2024.2321920.
5. Margaret M. Bradley, Mark K. Greenwald, Margaret C. Petry, and Peter J. Lang. Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):379–390, 1992. doi:10.1037/0278-7393.18.2.379.
6. Thackery I. Brown, Stephanie A. Gagnon, and Anthony D. Wagner. Stress disrupts human hippocampal-prefrontal function during prospective spatial navigation and hinders flexible behavior. *Current Biology*, 30(10):1821–1833.e8, May 2020. doi:10.1016/j.cub.2020.03.006.
7. Tad T. Brunyé, Caroline R. Mahoney, Jason S. Augustyn, and Holly A. Taylor. Emotional state and local versus global spatial memory. *Acta Psychologica*, 130(2):138–146, February 2009. doi:10.1016/j.actpsy.2008.11.002.
8. Sascha Credé, Tyler Thrash, Christoph Hölscher, and Sara Irina Fabrikant. The acquisition of survey knowledge for local and global landmark configurations under time pressure. *Spatial Cognition & Computation*, 19(3):190–219, July 2019. doi:10.1080/13875868.2019.1569016.
9. Rossana De Beni, Francesca Pazzaglia, Valérie Gyselinck, and Chiara Meneghetti. *Batteria per la valutazione delle abilità spaziali*. Giunti OS, Firenze, 2014.
10. Barbara L. Fredrickson. The role of positive emotions in positive psychology. *American Psychologist*, 56(3):218–226, July 2001. doi:10.1037/0003-066x.56.3.218.
11. Aaron Gardony and Holly A. Taylor. Affective states influence spatial cue utilization during navigation. *Presence: Teleoperators and Virtual Environments*, 20(3):223–240, June 2011. doi:10.1162/PRES_a_00046.
12. Carol A. Lawton. Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles*, 30(11-12):765–779, June 1994. doi:10.1007/BF01544230.
13. Veronica Muffato, Chiara Meneghetti, and Rossana De Beni. The role of visuo-spatial abilities in environment learning from maps and navigation over the adult lifespan. *British Journal of Psychology*, 111(1):70–91, February 2020. doi:10.1111/bjop.12384.
14. Veronica Muffato, Chiara Meneghetti, Virginia Di Ruocco, and Rossana De Beni. When young and older adults learn a map: The influence of individual visuo-spatial factors. *Learning and Individual Differences*, 53:114–121, January 2017. doi:10.1016/j.lindif.2016.12.002.
15. Alina Nazareth, Xun Huang, Daniel Voyer, and Nora S. Newcombe. A meta-analysis of sex differences in human navigation skills. *Psychonomic Bulletin & Review*, 26:1503–1528, 2019. doi:10.3758/s13423-019-01633-6.
16. Massimiliano Palmiero and Laura Piccardi. The role of emotional landmarks on topographical memory. *Frontiers in Psychology*, 8:763, May 2017. doi:10.3389/fpsyg.2017.00763.
17. Téophile Rasse, Valérie Gyselinck, and Jérôme Guegan. Encountering an emotional landmark on the route for a better spatial memory: What matters, valence or arousal? *Journal of Environmental Psychology*, 91:102145, November 2023. doi:10.1016/j.jenvp.2023.102145.
18. F. Ruotolo, M. H. G. Claessen, and I. J. M. Van Der Ham. Putting emotions in routes: the influence of emotionally laden landmarks on spatial memory. *Psychological Research*, 83(5):1083–1095, July 2019. doi:10.1007/s00426-018-1015-6.

19 Francesco Ruotolo, Filomena L. Sbordone, and Ineke J. M. Van Der Ham. The relationship between emotionally laden landmarks, spatial abilities, and personality traits: An exploratory study. *Brain Sciences*, 10(6):326, May 2020. doi:10.3390/brainsci10060326.

20 James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. doi:10.1037/h0077714.

21 Norbert Schwarz. Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440, July 2000. doi:10.1080/026999300402745.

22 Antonio Terracciano, Robert R. McCrae, and Paul T. Costa. Factorial and construct validity of the italian positive and negative affect schedule (panas). *European Journal of Psychological Assessment*, 19(2):131–141, April 2003. doi:10.1027/1015-5759.19.2.131.

23 Jan Wilkening and Sara Irina Fabrikant. The effect of gender and spatial abilities on map use preferences and performance in road selection tasks. In *Proceedings of the 25th International Cartographic Conference*, Paris, France, 2011. doi:10.5167/uzh-51316.

24 Jan Wilkening and Sara Irina Fabrikant. How do decision time and realism affect map-based decision making? In Max J. Egenhofer, editor, *Spatial Information Theory*, volume 6899 of *Lecture Notes in Computer Science*, pages 1–19. Springer, Berlin, Germany, 2011. doi:10.1007/978-3-642-23196-4_1.

25 Jan Wilkening and Sara Irina Fabrikant. How users interact with a 3d geo-browser under time pressure. *Cartography and Geographic Information Science*, 40(1):40–52, 2013. doi:10.1080/15230406.2013.762140.

26 Naoki Yamamoto and Giuseppe De Girolamo. Differential effects of aging on spatial learning through exploratory navigation and map reading. *Frontiers in Aging Neuroscience*, 4:14, 2012. doi:10.3389/fnagi.2012.00014.

Representing Retail Wayfinding: A Requirement-Driven Study Design for Non-Visual Supermarket Navigation

Xiang Zhou
Geoinformatics Group, Institute of Geography, University of Augsburg, Germany

Sabine Timpf
Geoinformatics Group, Institute of Geography, University of Augsburg, Germany

Abstract

First-visit supermarket shopping is a demanding case of non-visual indoor wayfinding. Blind and low-vision shoppers must make sense of functional zones, work through repetitive aisles, and confirm shelf areas while also dealing with queues and checkout. We use retail wayfinding to describe how movement, in-store search, and shelf-level confirmation are coupled. Retailer-held 2D layouts and product placement data can support verbal guidance, but many dynamic, sensory, and social conditions remain outside these data. We outline a requirement-driven mixed-methods study with 52 blind and low-vision participants: 30 in Study A, 30 in Study B, with eight participants taking part in both. Audio-recorded dialogue, verbal route descriptions, observer-coded event traces, and post-task interviews will be used to examine how shoppers request spatial information, confirm progress, and manage uncertainty from store-level orientation to shelf-level confirmation. The expected output is an initial coding matrix showing what needs to be represented, at what scale, and when confirmation or assistance becomes necessary in non-visual retail wayfinding.

2012 ACM Subject Classification Information systems → Spatial-temporal systems; Human-centered computing → Ubiquitous and mobile computing; Human-centered computing → Accessibility

Keywords and phrases retail wayfinding, blind and low-vision shoppers, indoor wayfinding, verbal route descriptions, requirements modelling

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.27

Category Short Paper

1 Introduction

Large indoor environments are difficult to navigate without vision because spatial information has to be selected, timed, and confirmed while the traveller is moving. Classic work describes non-visual wayfinding as problem solving around decision points, route knowledge, and uncertainty [18, 15], while guidance studies show that the form of spatial information affects cognitive load [7]. For blind and low-vision travellers, being on the correct path is therefore only part of the problem. Guidance also has to be understandable, confirmable, and able to support re-orientation when the perceived environment no longer matches the expected route [20, 6]. This paper focuses on first-visit supermarket shopping as a specific case of non-visual indoor wayfinding. A shopper enters with a list, moves between functional zones, searches within aisles, narrows down relevant shelf areas, and eventually reaches checkout.

As Figure 1 suggests, this recurrent task-chain can describe shopping more generally. The non-visual problem lies in how shoppers understand each stage, confirm progress, and re-establish orientation when expected cues are missing or unclear. Supermarkets make this harder because repetitive aisles and changing shelf displays make orientation unstable, while other customers affect where it is safe to stop, search, or ask for help.

© Xiang Zhou and Sabine Timpf;
licensed under Creative Commons License CC-BY 4.0

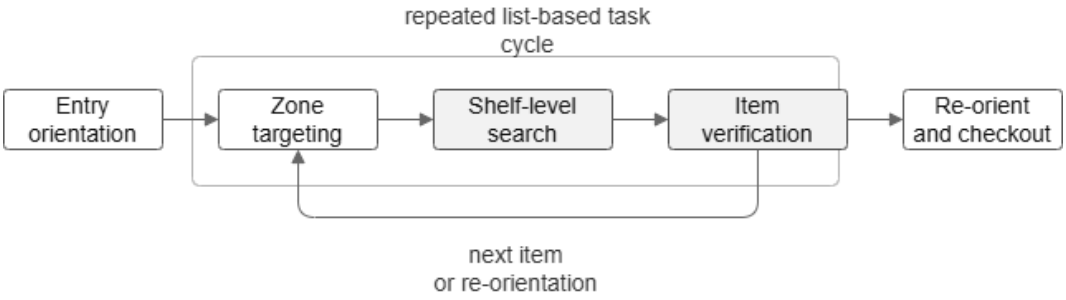


Figure 1 Recurrent shopping process and scale transitions during a first-visit supermarket task. The cycle is not unique to blind and low-vision shoppers; the paper focuses on the non-visual representation requirements that arise when this cycle has to be completed without visual access.

We use *retail wayfinding* to denote wayfinding in retail spaces where navigation is tightly coupled with in-store search and shelf-level confirmation. This definition builds on non-visual wayfinding and verbal route description research [18, 2], and responds to recent studies of grocery navigation, product descriptions, and shopping support for visually impaired people [14, 12, 3, 11]. In this setting, decision points are not only geometric locations such as aisle entrances. They also occur at functional moments in the shopping task: deciding whether an aisle is relevant, where to begin shelf-level search, or how to confirm that the relevant category area has been reached. Recent indoor navigation systems have made progress in positioning, semantic guidance, infrastructure-light interaction, and floor-plan-based support [21, 5, 10, 8]. Yet many still frame success as reaching a destination or point of interest. Supermarket guidance raises a more specific GIScience question: how can store structure and retailer-held spatial data be translated into information that supports non-visual shopping decisions?

Rather than presenting a completed navigation system, this paper asks which retail elements need to be represented, at which scale, and for which stage of a non-visual shopping task. We expect uncertainty to become especially visible at scale transitions, such as from store-level orientation to aisle-level travel, and from aisle-level travel to shelf-level confirmation. Some requirements may be represented from 2D layout and product placement data, while others require dynamic sensing, human knowledge, or interaction. The contribution is a requirements-focused study design with an initial coding matrix linking task stage, information need, spatial representation, and likely data source.

2 Background and Related Work

2.1 Wayfinding without vision

Research on wayfinding without vision shows that visually impaired travellers can acquire spatial knowledge, but the process depends strongly on how information is structured and timed. Passini and Proulx conceptualise wayfinding as problem solving around decision points, while Loomis, Klatzky, and Golledge connect non-visual navigation to route knowledge, landmarks, and spatial language [18, 15]. Montello's account of spatial knowledge acquisition is also useful here, because it distinguishes procedural route knowledge from broader configurational understanding [16]. In the orientation and mobility sense, support should therefore do more than compute a path: it should help the traveller maintain position, heading, relevant cues, upcoming decisions, and ways to re-establish orientation when an expected cue is not found [22, 23, 4].

2.2 Verbal route descriptions and non-visual guidance

Verbal route descriptions (VRDs) are valuable because they reveal how spatial information is selected and structured for action. They can include landmarks, distances, surfaces, sounds, object relations, and re-orientation cues. Balata's work on verbal route descriptions for non-visual navigation shows that such descriptions are richer than simple turn-by-turn commands and often contain information that supports orientation and confidence [2]. Recent work on dynamic indoor environments further suggests that verbal descriptions need to respond to movable objects, people, and changing local conditions [1].

In supermarkets, this matters because a useful instruction may need to explain why an aisle is relevant, where local search should begin, and which cue can confirm the right shelf segment. VRDs therefore offer not only a guidance modality, but also an empirical window into shopping-relevant spatial cognition.

2.3 Assistive indoor navigation and grocery shopping

Accessible indoor navigation systems have improved considerably, especially in positioning, semantic features, infrastructure-light interaction, and smartphone-based map interpretation [21, 5, 10, 8]. Studies of real-world use further show that visually impaired travellers adapt common navigation technologies to situated needs, rather than simply following pre-computed routes [6]. Yet many systems still treat success mainly as reaching a destination or a point of interest. This framing is only partly adequate for supermarket shopping. A shopper may reach the correct department and still struggle at the shelf, because category boundaries, shelf organisation, prices, and packaging cues remain uncertain.

Recent work on grocery navigation, product descriptions, and wearable shopping assistance addresses this gap from complementary directions [14, 12, 3]. Grocery navigation studies show how mobile guidance can support store movement, product-description work clarifies what information is needed to distinguish similar items, and wearable assistance highlights the difficulty of shelf-level recognition in dense scenes. The last-few-metres problem is relevant because uncertainty often appears close to the target, where local confirmation and object interaction become necessary [20]. What remains open is how blind and low-vision shoppers connect store layout, functional zones, sensory cues, and product search during an unfamiliar trip. We address this as a representation problem rather than another positioning or product-recognition system.

3 Retail Wayfinding as a Distinct Problem

Supermarket retail wayfinding is not reducible to indoor wayfinding plus product identification. Instead of following one route to one destination, the shopper completes a list through repeated cycles of moving, searching, confirming, and re-orienting. Reaching the right zone is useful, but it is not enough if the shopper cannot decide where to begin searching, how to confirm the relevant category or shelf area, or how to re-establish orientation when the expected shelf arrangement is not found [14, 12, 3]. Retail wayfinding is therefore used here as a spatial representation problem, not just a task label. The difficulty comes from doing list-based shopping in repetitive shelf spaces, where product categories can be ambiguous and conditions change with carts, displays, restocking, and crowding [17].

Retailer-held spatial data make this distinction concrete. A 2D floor plan can represent entrances, checkout areas, aisles, cross-aisles, and broad functional zones, while product placement data indicate where categories or items are expected to be. Together, these data

can support store overviews, zone-level routing, and some aisle-level guidance. However, they rarely capture shelf-level accessibility, temporary obstacles, crowding, sound masking, or local confirmation cues. The research problem is therefore to determine which parts of the shopping task can be represented from stable spatial data, and where non-visual guidance needs sensing, human knowledge, or interaction.

4 A Requirements-Oriented Framework for Retail Wayfinding

We use a requirements-oriented framework to translate supermarket shopping episodes into spatial representation requirements. It links shoppers' needs to supermarket structure and verbal guidance implications, while keeping the focus on representation rather than system architecture. When a shopper asks a question, hesitates, requests confirmation, or reconstructs a route after the task, this mapping helps identify what spatial knowledge was needed and whether it could plausibly be derived from retailer-held data. The framework is an initial coding lens to be refined through the empirical study.

Supermarkets are treated as functional spatial systems, from entrances and checkout areas to aisles, shelf segments, and points where shoppers confirm progress. These elements operate across store, zone, aisle, and shelf scales. Uncertainty often increases when shoppers move from travelling to searching or from searching to confirming.

Table 1 summarises this mapping as an initial coding matrix. The four requirements provide a working lens derived from prior work on non-visual wayfinding, verbal guidance, cognitive load, in-situ navigation technology, last-few-metres wayfinding, and remote sighted assistance [18, 7, 2, 6, 20, 13]. In the empirical study, the table will be refined into a task-stage by need matrix linking each episode to the spatial representation required at that moment.

5 Planned Study and Coding Approach

Subject to ethical approval, the study will be conducted in June–July 2026 in unfamiliar supermarkets with 52 blind and low-vision adults. Study A will involve 30 participants in an in-situ shopping task, while Study B will involve 30 participants in a complementary verbal description and post-task reconstruction task; eight participants will take part in both. This partial overlap connects the two components without treating them as one repeated-measures experiment.

In Study A, participants will locate predefined product categories from different functional zones and then move toward checkout. The store will be checked in advance only to ensure that these target categories are present and can be encountered in plausible shelf areas; no exact product choice is prescribed. A human guide will act as a proxy for future non-visual guidance, prioritising participant requests and safety-critical interventions over unsolicited route planning. The guide may confirm that a participant has reached the relevant category area or intervene for safety, but will not identify a specific product as the correct purchase. Guide-initiated statements and participant-initiated requests will be coded separately. The aim is not to test product recognition or exact item selection, but to identify what spatial and semantic information becomes necessary from store-level orientation to category-level search and shelf-level confirmation.

The study will combine audio-recorded in-situ dialogue with verbal route descriptions, observer-coded event traces, and short post-task interviews. Event traces will be logged as time-stamped episodes, covering the main transition points from entry orientation to checkout. These traces do not evaluate route optimality; instead, they show where participants moved

Table 1 Initial coding matrix linking non-visual needs, spatial knowledge, representational basis, and verbal guidance implications in supermarket shopping.

Need	Spatial knowledge needed	Representational basis / likely data source	Verbal guidance implications
Safety	Safe corridors, reliable boundaries, safe stopping points near search locations [18, 7, 6]	2D geometry + optional dynamic updates	Edge-following cues, clearance cues, safe pause points before search [2, 6, 14]
Control	Preferred pacing; available clarifications per task stage [9, 6, 13]	Task-stage model + dialogue state + user preferences	User-driven queries; progressive disclosure; repeat/rewind; transition cues for local search [9, 6, 13]
Understanding	Zones and relations; anchors relevant to shopping (checkout, entrance, service counter) [16, 21, 8]	2D plan + semantic zoning	Short store overview; anchor references; boundary references between adjacent zones [21, 8, 19]
Confidence and re-orientation	Confirmation cues at aisle entrances, shelf ends, and category boundaries under aisle ambiguity [20, 2, 1, 14]	Topology + decision-point model + cue hierarchy from VRDs	Confirmation checks; conditional return cues; aisle-identity confirmation [20, 2, 1, 13]

smoothly and where uncertainty or assistance became visible. Interviews will ask participants to reconstruct the trip and identify trusted cues, moments of uncertainty, and situations where assistance became necessary.

The analytical unit will be an event episode: a bounded segment of dialogue or movement around a task-relevant spatial action. Each episode will be coded by task stage, need type, cue type, spatial scale, decision context, assistance source, and likely data source. Task stages include entry orientation, zone targeting, aisle traversal, shelf-level search, shelf-area confirmation, re-orientation, and checkout. Cue types include structural, semantic, tactile, auditory, olfactory, and social cues. Decision contexts distinguish geometric decision points from functional ones, such as choosing where to begin a shelf search. The resulting matrix will compare what participants asked for in situ, what they reconstructed after the task, and which environmental features were represented by 2D layout or product placement data. The main analytical task is to move from recorded interaction to representation requirements: identifying stable coding categories, overlaps, and cases where sensing, human knowledge, or interaction remains necessary.

6 Conclusion

Retail wayfinding frames first-visit supermarket shopping as more than destination-based indoor navigation. The shopping cycle itself is not unique to blind and low-vision shoppers. What is specific is the need to represent movement, in-store search, shelf-level confirmation, and re-orientation in forms that can be understood and confirmed without visual access. The framework developed here relates shoppers' needs to supermarket structure and to the

limits of retailer-held 2D layout and product placement data. The planned mixed-methods study with 52 participants provides the empirical basis for refining this framework. Its main analytical task is to turn the recorded dialogue, route descriptions, event traces, and post-task reconstruction into a coding matrix for task stages, spatial scale, assistance, and confirmation. This matrix makes the data comparable across what participants request in situ, what they remember after the task, what can be represented from stable retailer-held data, and what depends on dynamic sensing or human knowledge. By focusing on this transition from observed shopping behaviour to representation requirements, the paper supports later work on verbal guidance, spatial data modelling, and non-visual supermarket navigation support.

References

1 Abdulaziz Alrashidi, Peter Cudd, Charith Abhayaratne, and Yoshihiko Gotoh. Exploration of verbal descriptions and dynamic indoors environments for people with sight loss. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 110:1–110:6. ACM, 2023. doi:10.1145/3544549.3585883.

2 Jan Balata. *Verbal Route Descriptions in Navigation of People with Visual Impairments*. PhD thesis, Czech Technical University, August 2018.

3 Roger Boldu, Denys J. C. Matthies, Haimo Zhang, and Suranga Nanayakkara. AiSee: An assistive wearable device to support visually impaired grocery shoppers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4):119:1–119:25, 2020. doi:10.1145/3432196.

4 Nicholas A. Giudice and Richard G. Long. Establishing and maintaining orientation: Tools, techniques, and technologies. In Robert S. Wall Emerson, Bruce B. Blasch, and William R. Wiener, editors, *Foundations of Orientation and Mobility: Vol. I, History and Theory*, pages 18–54. APH Press, 4 edition, 2024. URL: https://www.researchgate.net/publication/385852439_Establishing_and_Maintaining_Orientation_Tools_Techniques_and_Technologies.

5 Nicholas A Giudice, William E Whalen, Timothy H Riehle, Shane M Anderson, and Stacy A Doore. Evaluation of an accessible, real-time, and infrastructure-free indoor navigation system by users who are blind in the Mall of America. *Journal of Visual Impairment & Blindness*, 113(2):140–155, 2019. doi:10.1177/0145482X19840918.

6 Vaishnav Kameswaran, Alexander J. Fiannaca, Melanie Kneisel, Amy Karlson, Edward Cutrell, and Meredith Ringel Morris. Understanding in-situ use of commonly available navigation technologies by people with visual impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 28:1–28:12. ACM, 2020. doi:10.1145/3373625.3416995.

7 Roberta L Klatzky, James R Marston, Nicholas A Giudice, Reginald G Golledge, and Jack M Loomis. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied*, 12(4):223–232, 2006. doi:10.1037/1076-898X.12.4.223.

8 Masaya Kubota, Masaki Kuribayashi, Seita Kayukawa, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. Snap&Nav: Smartphone-based indoor navigation system for blind people via floor map analysis and intersection detection. *Proc. ACM Hum. Comput. Interact.*, 8(MHCI):1–22, 2024. doi:10.1145/3676522.

9 Bineeth Kuriakose, Ida Marie Ness, Maja Åskov Tengstedt, Jannicke Merete Svendsen, Terese Bjørseth, Bijay Lal Pradhan, and Raju Shrestha. *Turn Left Turn Right* - delving type and modality of instructions in navigation assistant systems for people with visual impairments. *Int. J. Hum. Comput. Stud.*, 179:103098, 2023. doi:10.1016/j.ijhcs.2023.103098.

10 Masaki Kuribayashi, Tatsuya Ishihara, Daisuke Sato, Jayakorn Vongkulbhisal, Karnik Ram, Seita Kayukawa, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. PathFinder: Designing a map-less navigation system for blind people in unfamiliar buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 41:1–41:16. ACM, 2023. doi:10.1145/3544548.3580687.

11 Jihyun Lee, Jinsol Kim, and Hyunggu Jung. Challenges and design opportunities for easy, economical, and accessible offline shoppers with visual impairments. In *Proceedings of AsianCHI 2020*, pages 69–72. ACM, 2020. doi:10.1145/3391203.3391223.

12 Kyungyeon Lee, Sohyeon Park, and Uran Oh. Designing product descriptions for supporting independent grocery shopping of people with visual impairments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 425:1–425:6. ACM, 2021. doi:10.1145/3411763.3451806.

13 Sooyeon Lee, Madison Reddie, Chun-Hua Tsai, Jordan Beck, Mary Beth Rosson, and John M. Carroll. The emerging professional practice of remote sighted assistance for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2020. doi:10.1145/3313831.3376591.

14 Chengpu Liao, Xingpan Jin, and EunJeong Cheon. EasyGO: A field study of grocery store navigation application design for the visually impaired. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference*. ACM, 2024. doi:10.1145/3656156.3663719.

15 Jack M Loomis, Roberta L Klatzky, Reginald G Golledge, et al. Navigating without vision: basic and applied research. *Optometry and Vision Science*, 78(5):282–289, 2001. doi:10.1097/00006324-200105000-00011.

16 Daniel R Montello. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. *Spatial and temporal reasoning in geographic information systems*, pages 143–154, 1998. doi:10.1093/oso/9780195103427.003.0011.

17 Soobeen Park and Shaoqing Zhang. A pilot study of circulation layout based on perceived retail crowding. *Journal of Retailing and Consumer Services*, 49:305–315, 2019. doi:10.1016/j.jretconser.2019.04.008.

18 Romedi Passini and Guyltne Proulx. Wayfinding without vision: An experiment with congenitally totally blind people. *Environment and Behavior*, 20(2):227–252, 1988. doi:10.1177/0013916588202006.

19 Sebastian Rottmann, Claudia Loitsch, and Gerhard Weber. Accessible mobile map application and interaction for people with visual or mobility impairments. In *Proceedings of Mensch und Computer 2022*, pages 119–127. ACM, 2022. doi:10.1145/3543758.3543780.

20 Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. Closing the gap: Designing for the last-few-meters wayfinding problem for people with visual impairments. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 222–235. ACM, 2019. doi:10.1145/3308561.3353776.

21 Daisuke Sato, Uran Oh, João Guerreiro, Dragan Ahmetovic, Kakuya Naito, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. NavCog3 in the wild: Large-scale blind indoor navigation assistant with semantic features. *ACM Trans. Access. Comput.*, 12(3):14:1–14:30, 2019. doi:10.1145/3340319.

22 Victor R Schinazi, Tyler Thrash, and Daniel-Robert Chebat. Spatial navigation by congenitally blind individuals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(1):37–58, 2016. doi:10.1002/wcs.1375.

23 Catherine Thinus-Blanc and Florence Gaunet. Representation of space in blind persons: vision as a spatial sense? *Psychological bulletin*, 121(1):20, 1997. doi:10.1037/0033-2909.121.1.20.

Using Social Cues for Wayfinding: An Eye-Tracking Study of Attention Allocation and Evidence Accumulation in a Virtual Reality T-Intersection

Yu Qiao
School of Geography, University of Leeds, UK

Carlo Campagnoli
School of Psychology, University of Leeds, UK

Ed Manley[1]
School of Geography, University of Leeds, UK

Abstract

In synchronous weak social wayfinding, route decisions are shaped in real time by others' behaviour, yet the sampling and integration of social information during decision-making remain poorly understood. We investigated whether attention to pedestrians in a virtual-reality T-intersection reflects evidence accumulation preceding route commitment. Participants navigated scenes containing avatar crowds in which crowd size, proportion of credible others, directional majority, walking speed, and movement style were manipulated in an incomplete within-subjects orthogonal design. Continuous eye tracking measured attention to social cues, the destination landmark, and the surrounding scene. Results showed that gaze was directed predominantly toward social cues, whereas direct inspection of the landmark was rare. Among the manipulated factors, movement style and walking speed produced the clearest, though small, effects on fixation–saccade dynamics. Crucially, social gaze rose sharply early in the trial and then declined gradually before route commitment, suggesting that participants initially sampled social evidence and progressively reduced inspection as the decision unfolded. These findings indicate that social navigation involves accumulating and integrating others' behaviour over time rather than simply following them.

2012 ACM Subject Classification Information systems → Spatial-temporal systems; Human-centered computing → Empirical studies in HCI; Applied computing → Psychology

Keywords and phrases Social navigation, wayfinding, eye tracking, virtual reality, evidence accumulation, attention

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.28

Category Short Paper

1 Introduction

Wayfinding is not merely a geometric problem of choosing the shortest or most direct path. In naturalistic environments, route decisions are shaped by spatial layout, landmark structure, and the behaviour of other people [8, 3, 14, 2]. Dalton et al. [4] describe wayfinding as a social activity and distinguish weak synchronous social wayfinding, in which route choice is shaped in real time by others' behaviour without direct interaction, from stronger or temporally separated forms. Evidence from virtual environments likewise shows that crowd density and movement can systematically alter route choice and locomotion [11, 6, 7].

A key question in social wayfinding is how information from other people influences route decisions. For social information to affect behaviour, it must first be attended to. Attention does not merely accompany choice; it determines which information is sampled and how it is

[1] corresponding author

© Yu Qiao, Carlo Campagnoli, and Ed Manley;
licensed under Creative Commons License CC-BY 4.0

weighted during deliberation, shaping evidence accumulation and the eventual decision [9, 12]. Work on eye movements in decision making further shows that gaze can actively modulate information processing rather than passively reflect visual interest [10, 13]. This perspective suggests that gaze can provide insight into how social information is gathered and used during navigation. We therefore ask three related questions: how gaze is allocated between social cues and spatial cues, and among different social cues; how social gaze unfolds over time within a trial; and whether the temporal dynamics of social gaze are consistent with a process of evidence accumulation.

Eye tracking is therefore well suited to testing the process-based account of social navigation, especially in virtual reality, where route choice can be studied under controlled but behaviourally meaningful conditions. Prior work suggests that VR can reproduce key aspects of pedestrian route choice [5]. Yet challenges remain: fixation-saccade classification is less straightforward in dynamic 3D scenes [1], and gaze-to-object mappings are affected by geometric bias, because nearer objects are more likely to intersect gaze rays and may therefore appear more salient if uncorrected.

Against this background, the present study examined route choice in a virtual reality T-intersection. Participants viewed crowds of avatars moving toward one branch or the other under experimental control, while continuous eye tracking measured attention to social cues, the destination landmark, and the broader scene. The aim was to characterize how attention is allocated to social information during route choice and how social gaze evolves over time. In particular, we examined whether the temporal dynamics of social gaze are consistent with a process of evidence accumulation during navigation. This design captures both overall attention allocation and the moment-to-moment evolution of gaze prior to behavioural commitment.

This study makes three contributions. First, it extends social wayfinding research by treating gaze toward pedestrians as an evidentiary process unfolding before route commitment. Second, it improves methodology by combining sensitivity analysis for fixation-saccade classification with a geometric correction for distance-related hit probability. Third, it clarifies how social cues compete with spatial information in a controlled T-intersection and whether avatar motion modulates attention in a manner consistent with staged evidence accumulation. Together, these contributions position social gaze as a process-level bridge between perception and route decision in virtual navigation.

2 Methods

2.1 Experimental design and procedure

The experiment used an incomplete within-subjects orthogonal design based on an L36 array to manipulate five social-cue factors: pedestrian count, proportion of credible others, directional majority, walking speed, and movement style. Each participant completed nine unique conditions sampled from the full factorial set, with no repeated condition within participant. Trial order was randomized individually to reduce sequence effects.

The virtual environment was a minimal T-shaped intersection with identical geometry on both branches (Figure 1). The task was a route-decision task: at each trial, participants were instructed to imagine that they were on a city street, that they were trying to reach the stadium to watch a football match, and that they did not know which direction would lead to the destination. They were asked to use the available social and spatial cues to decide whether to go left or right. The destination landmark, operationalized as a stadium, was

Table 1 Experimental variables and levels.

Variable	Target path	Non-target path
Pedestrian count	Total pedestrians: 5, 20, 40	
Proportion of credible others	0	0, 20%, 75%
Directional majority	45%, 25%, 10%	55%, 75%, 90%
Walking speed	1.4 m/s	0.6, 1.4, 2.2 m/s
Movement style	Direct	Direct, Uncertain

placed behind one branch (target path) and counterbalanced across left and right sides across trials. This kept the destination available as a navigational reference without making one branch trivially dominant.

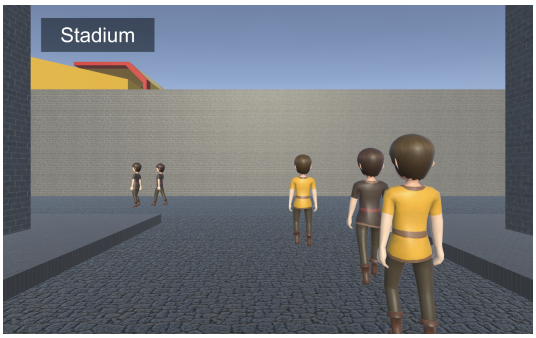


Figure 1 Virtual T-intersection environment.

Avatars were spawned along the approach corridor, reached the intersection at controlled timings, and then selected a branch according to the directional-majority manipulation before continuing with the assigned speed and movement style. The manipulations are summarized in Table 1. Pedestrian count varied across low, medium, and high levels. Credible others were operationalized as the percentage of avatars wearing football jerseys, which served as an association cue in this study. Directional majority indexed the extent to which the crowd favoured one branch. Walking speed varied across slow, intermediate, and fast levels. Movement style contrasted direct, goal-oriented motion with uncertain or wavering motion. The dependent variable was the participant's final route choice, recorded after 40 seconds when a response panel appeared and participants selected either the left or right branch.

A priori power analysis ($\alpha = .05$, power $= .80$) indicated that $N = 61$ was sufficient to detect medium-sized main effects (Cohen's $f \approx 0.25$) and typical interaction effects under the incomplete within-subjects L36 design. A total of 61 participants were recruited from the university participant pool and the local community. All participants had normal or corrected-to-normal vision and reported no severe motion or vestibular disorders. The study was approved by the Research Ethics Committee for Business, Environment, and Social Sciences at the University of Leeds, and all participants provided informed consent before participation. Participants received either a voucher or course credit as compensation.

2.2 Eye-movement classification and sensitivity analysis

Because the raw gaze log did not contain event-level labels, fixations and saccades were classified post hoc using an angular-velocity criterion. For each trial, gaze direction was represented as a unit vector at each sample. Angular displacement between consecutive

samples was converted to angular velocity (deg/s), and samples exceeding a predefined threshold were labelled as saccades, with the remainder labelled as fixations. This yielded a trial-wise eye-state series aligned with areas of interest for subsequent analyses.

To determine an appropriate threshold, a sensitivity analysis was conducted over candidate values ranging from 2 to 100 deg/s. For each threshold, the full analysis pipeline was rerun and evaluated using four criteria: fixation–saccade balance, availability of valid time bins for social-attention summaries, magnitude of the balance index, and proportion of usable bins across factors. These were combined into a composite score favouring thresholds that preserved temporal variation while maintaining adequate coverage and moderate saccade rates. A threshold of 10 deg/s provided the optimal trade-off (saccade rate: 15.3%; coverage: 93.6%), and neighbouring values (7.5 and 12.5 deg/s) produced comparable results, indicating robustness. Accordingly, 10 deg/s was adopted for all analyses.

2.3 Geometric correction for distance-related hit probability

A further challenge in VR eye tracking is geometric sampling bias: nearer objects subtend larger visual angles and are therefore more likely to intersect a gaze ray independent of attentional relevance. To mitigate this effect, a condition-specific geometric baseline was estimated using Monte Carlo simulations of avatar trajectories.

For each condition, simulations reproduced the same spawning rules, movement parameters, and temporal structure as the VR task. At each time point, the probability that avatar i would be intersected by a gaze ray was approximated as:

$$p_i(t) = \begin{cases} 0, & z_i(t) \leq 0 \\ \min\left(1, \frac{A}{d_i(t)^2}\right), & z_i(t) > 0 \end{cases} \tag{1}$$

where $d_i(t)$ denotes the Euclidean distance to the observer, A is the effective projected frontal area of a pedestrian proxy, and $z_i(t)$ indicates depth relative to the camera plane. Probabilities were aggregated across avatars and averaged over time to obtain trial-level visibility, then averaged across simulation runs to derive condition-specific baselines. Empirical gaze measures were normalised against these baselines, allowing attention to be interpreted relative to expected geometric visibility.

Although this approach does not fully account for all visibility factors (e.g., occlusion), it substantially reduces distance-related bias and improves the validity of gaze-based attention measures.

3 Results

3.1 Overall allocation of attention

Across observations, gaze was directed mainly to social cues and other scene elements, whereas the spatial cue (the stadium) received minimal direct inspection. Social areas of interest accounted for a mean of 35.6% of gaze samples, other areas for 64.2%, and the spatial cue for only 0.2%. The median and upper quartile of spatial gaze were both 0.0%, indicating that direct fixations on the landmark were absent in most trials. In contrast, social gaze varied substantially across trials, suggesting that participants allocated more visual attention to pedestrian information than to the destination landmark within the present experimental setup.

This pattern was also evident at the condition level (Figure 2, left). Across the 36 experimental conditions, social gaze ranged from approximately 14% to 64%, with higher social allocation corresponding to lower gaze to other scene elements. Spatial gaze remained negligible in nearly all conditions and did not materially alter this ranking. Thus, the manipulations redistributed attention primarily between social and non-social scene elements rather than producing substantial inspection of the landmark. The participant-level plot showed substantial inter-individual variation, with some participants allocating more than 70% of gaze to social cues and others less than 20% (Figure 2, right). Even so, the same general structure held: social gaze and other gaze formed a near-complementary distribution, whereas spatial gaze remained marginal.

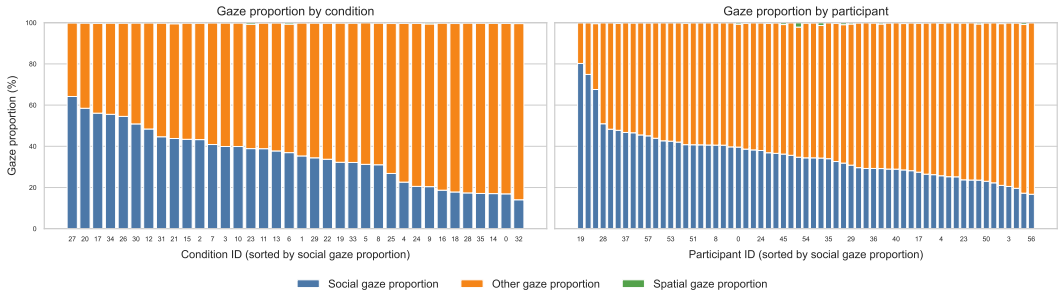


Figure 2 Gaze proportion by condition (left) and by participant (right).

3.2 Fixation–saccade dynamics and social cue properties

Using the 10 deg/s threshold, the gaze stream was strongly fixation-dominant: 84.7% of samples were classified as fixations and 15.3% as saccades. This indicates that participants spent most of the task in stable viewing states, with saccades functioning mainly as brief transitions between informative scene elements. The factor-level analyses therefore examined whether specific social cues altered this balance over time.

Across the three social factors, level-by-state contingency tests were significant in all cases, but effect sizes were small. Outfit showed a weak association with eye state ($\chi^2(1) = 58.98$, $p = 1.59 \times 10^{-14}$, Cramér's $V = 0.0068$), movement style showed the strongest association ($\chi^2(1) = 318.77$, $p = 2.68 \times 10^{-71}$, Cramér's $V = 0.0158$), and walking speed also reached significance ($\chi^2(2) = 104.31$, $p = 2.24 \times 10^{-23}$, Cramér's $V = 0.0090$). These results suggest that social cues biased visual sampling without producing large categorical changes in gaze behaviour. In other words, the cues did not determine whether participants looked, but rather how long they remained in a stable sampling state before shifting to another source of information.

The time-resolved balance maps provided a clearer picture of these effects. Outfit produced only limited differences between levels, with both conditions remaining fixation-dominant throughout most of the trial (Figure 3, top). Movement style showed the clearest temporal contrast: uncertain pedestrians were associated with a more fixation-weighted profile, whereas direct motion produced relatively more saccades, especially later in the trial (Figure 3, middle). Walking speed also modulated gaze allocation in a graded manner, with slower motion producing stronger fixation dominance and faster motion leading to earlier changes in balance (Figure 3, bottom). Overall, movement style had the strongest effect, speed a weaker but structured influence, and outfit the smallest impact.

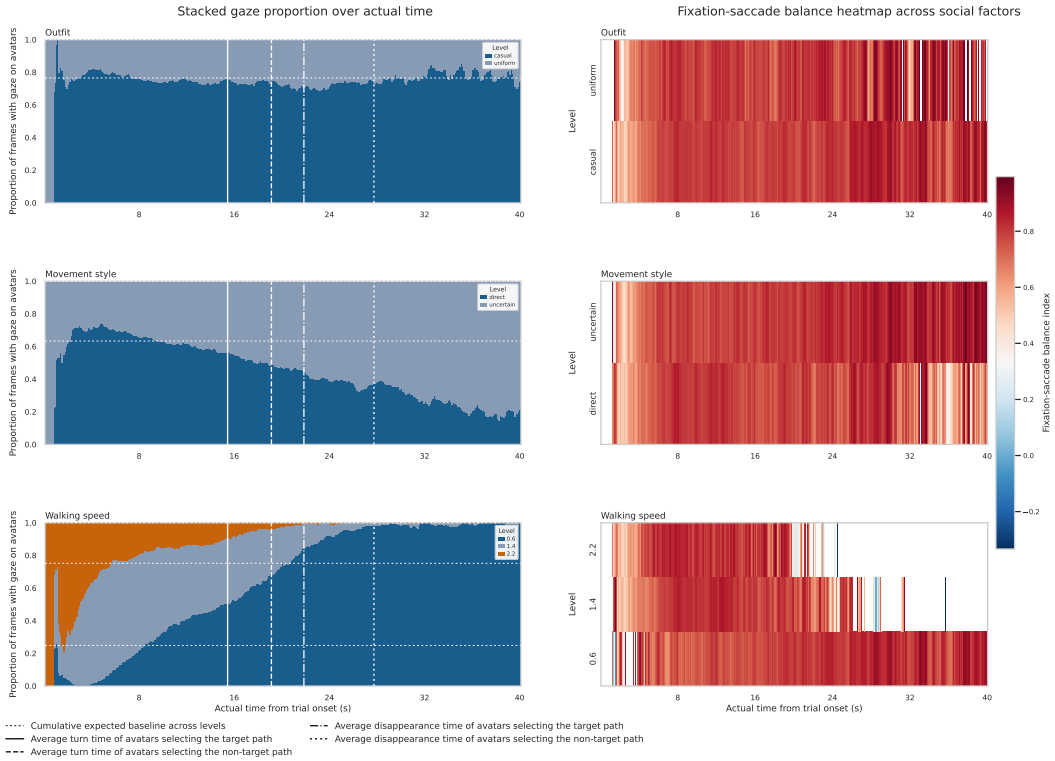


Figure 3 Left: Stacked gaze proportions over time across social factors (outfit, movement style, and walking speed; top to bottom). Right: Fixation–saccade balance heatmaps across social factors (outfit, movement style, and walking speed; top to bottom), where the balance index is defined as (Fixation − Saccade) / (Fixation + Saccade).

3.3 Temporal dynamics of social gaze as evidence accumulation

The temporal profile of social gaze suggests an evidence-accumulation process. In the raw data, mean social gaze rose sharply from near zero after trial onset, peaked at around 8–10 s, and then declined gradually for the rest of the trial (Figure 4, left). This early surge suggests rapid sampling of pedestrian information to form an initial directional hypothesis, followed by progressive redistribution of gaze as the decision advanced. Importantly, this peak occurred before the main behavioural commitments of the avatars, indicating that substantial evidence accumulation took place prior to turning.

After geometric correction, the same temporal structure remained but became more interpretable (Figure 4, right). The corrected curve preserved the early rise and peak, although its magnitude was reduced, indicating that part of the raw effect reflected distance-related hit probability rather than pure attentional preference. Once this bias was removed, the trajectory appeared less uniformly elevated and more clearly divided into stages of accumulation, consolidation, and stabilization. The turn of avatars selecting the target path occurred shortly after the peak of social attention, suggesting that participants likely reached sufficient evidence to initiate commitment while still monitoring social cues. After the relevant avatars left the scene, the corrected curve settled into a lower and more stable baseline with only small late-trial rebounds.

Overall, the temporal dynamics support a staged account of decision-making in which attention first accumulates social evidence, then tracks the commitment point, and finally stabilizes after the critical cues disappear. The gaze trajectory is therefore not only descriptive of where participants looked, but also informative about how route-relevant evidence was sampled, evaluated, and converted into a decision.

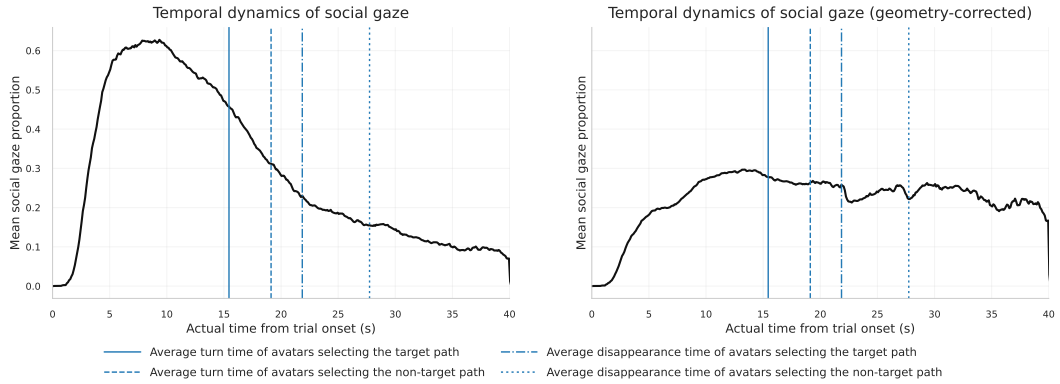


Figure 4 Temporal dynamics of social gaze before (left) and after (right) geometric correction.

4 Discussion

This study examined whether social information in wayfinding can be understood as an evidence-accumulation process. The results support this interpretation. Participants directed most of their gaze to social cues rather than to the destination landmark, indicating that, in this experimental context, the crowd was the main source of route decision-relevant information. However, this pattern should be interpreted within the present experimental context and does not imply that social cues are generally more important than spatial cues in naturalistic wayfinding. Social gaze also showed a clear temporal pattern: it rose rapidly at trial onset, peaked before or around the main turning window, and then declined as the scene unfolded. This trajectory is consistent with early evidence gathering followed by progressive commitment to a route.

Geometric correction further strengthened this account. In the raw data, social gaze was partly inflated by proximity-based hit probability, because nearby avatars were easier to intersect with a gaze ray. After correction, the same temporal structure remained but became more clearly linked to decision timing rather than spatial bias. This suggests that social gaze reflected not only visual accessibility, but also the informational value of pedestrian behaviour at different moments.

The fixation-saccade results are also consistent with a process account. Fixations dominated overall, indicating that participants spent most of the task in stable sampling states, while saccades functioned mainly as transitions between informative scene elements. Among the social cues, movement style was the most informative, with uncertain motion encouraging longer inspection. Walking speed showed a similar but weaker effect, suggesting that dynamic cues modulate the rate of evidence sampling rather than the mere presence of social attention. The small effect sizes should therefore not be taken as null effects; instead, they suggest that route choice was shaped by multiple weak cues integrated over time rather than by any single all-or-none trigger. This interpretation also points to a natural extension of the

present work: a drift diffusion model with time-varying social evidence, in which gaze-gated avatar information is translated into a momentary evidence stream and accumulated over time to predict route choice more directly.

Methodologically, the study highlights the value of combining VR with careful gaze analysis. Threshold-based fixation classification is necessary in this setting, but threshold choice should be guided by sensitivity analysis rather than convention. Geometric correction is also essential when objects differ in distance and projected size, because apparent social attention can otherwise be overestimated. Future work should extend the correction model to include occlusion and line-of-sight constraints, which also shape visibility in crowd scenes.

The study has limitations. The environment was simplified to isolate social information, which improves internal validity but reduces ecological complexity. More importantly, route choice was reported only at the end of each 40-second trial, and explicit decision times were not collected, so some gaze samples may have been recorded after participants had already made their decision. In addition, gaze to avatars may reflect general attentional bias toward humans, motion saliency, or rendering-specific visual properties, rather than navigation-specific sampling of social cues. Future studies should record continuous choice timing, which would make it possible to test dynamic accumulation models more directly and to separate perceptual sampling from the final commitment to a route, and examine whether the proposed accumulation account generalizes to more complex and interactive urban environments.

5 Conclusion

This study shows that social information can guide wayfinding through evidence accumulation. Participants relied on pedestrians far more than on the destination landmark, and their gaze followed a clear temporal pattern: rapid early sampling, a peak before route commitment, and a later decline as the decision resolved. Movement style and walking speed had the clearest effects, but these were graded rather than dramatic. After correcting for geometric visibility, the temporal dynamics became especially compelling as evidence for a process account.

More broadly, social navigation appears to be a dynamic decision process in which people gather, compare, and integrate social evidence until it becomes sufficient for route choice. Future work could formalize this process with computational models of evidence accumulation, such as drift–diffusion models, combined with time-varying social cues derived from gaze behaviour. This would provide a more mechanistic account of social navigation and link observed gaze dynamics to latent decision processes.

References

1 Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye Tracking in Virtual Reality: A Broad Review of Applications and Challenges. *Virtual reality*, 27(2):1481–1505, 2023. doi:10.1007/s10055-022-00738-z.

2 Madeleine Alcock, Jan M. Wiener, and Doug Hardman. Searching in an unfamiliar environment: A phenomenologically informed experiment. *Phenomenology and the Cognitive Sciences*, 2024. doi:10.1007/s11097-024-09960-3.

3 Debjit Bhowmick, Stephan Winter, Mark Stevenson, and Peter Vortisch. The impact of urban road network morphology on pedestrian wayfinding behavior. *Journal of Spatial Information Science*, 21:203–228, 2020. doi:10.5311/JOSIS.2020.21.601.

4 Ruth C. Dalton, Christoph Hölscher, and Daniel R. Montello. Wayfinding as a Social Activity. *Frontiers in Psychology*, 10, 2019. doi:10.3389/fpsyg.2019.00142.

5 Yan Feng, Dorine C. Duives, and Serge P. Hoogendoorn. Using virtual reality to study pedestrian exit choice behaviour during evacuations. *Safety Science*, 137:105158, 2021. doi:10.1016/j.ssci.2021.105158.

6 Milad Haghani and Majid Sarvi. Pedestrian crowd tactical-level decision making during emergency evacuations. *Journal of Advanced Transportation*, 50(8):1870–1895, 2016. doi:10.1002/atr.1434.

7 Hengshan Li, Tyler Thrash, Christoph Hölscher, and Victor R. Schinazi. The effect of crowdedness on human wayfinding and locomotion in a multi-level virtual shopping mall. *Journal of Environmental Psychology*, 65:101320, 2019. doi:10.1016/j.jenvp.2019.101320.

8 Rui Li and Alexander Klippel. Wayfinding Behaviors in Complex Buildings: The Impact of Environmental Legibility and Familiarity. *Environment and Behavior*, 48(3):482–510, 2016. doi:10.1177/0013916514550243.

9 Felix Molter, Armin W. Thomas, Scott A. Huettel, Hauke R. Heekeren, and Peter N. C. Mohr. Gaze-dependent evidence accumulation predicts multi-alternative risky choice behaviour. *PLoS Computational Biology*, 18(7):e1010283, 2022. doi:10.1371/journal.pcbi.1010283.

10 Jacob L. Orquin and Simone Mueller Loose. Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, 144(1):190–206, 2013. doi:10.1016/j.actpsy.2013.06.003.

11 Carlota Silva, Francisco Rebelo, Elisângela Vilar, and Paulo Noriega. Preliminary Study about Social Influence Over Wayfinding Decisions. *Procedia Manufacturing*, 3:5920–5926, 2015. doi:10.1016/j.promfg.2015.07.905.

12 Chih-Chung Ting and Sebastian Gluth. Unraveling information processes of decision-making with eye-tracking data. *Frontiers in Behavioral Economics*, 3, 2024. doi:10.3389/frbhe.2024.1384713.

13 Michel Wedel, Rik Pieters, and Ralf van der Lans. Modeling Eye Movements During Decision Making: A Review. *Psychometrika*, 88(2):697–729, 2023. doi:10.1007/s11336-022-09876-4.

14 Demet Yesiltepe, Ruth Conroy Dalton, and Ayse Ozbil Torun. Landmarks in wayfinding: A review of the existing literature. *Cognitive Processing*, 22(3):369–410, 2021. doi:10.1007/s10339-021-01012-x.

Feeling the Heat? Heat Stress Impairs Landmark but Not Route Knowledge in a VR Navigation Task

Alessandro J. Pierro
Independent Scholar, Zurich, Switzerland

Sara I. Fabrikant
Department of Geography and Digital Society Initiative, University of Zurich, Switzerland

Armand Kapaj[1]
Department of Geography and Digital Society Initiative, University of Zurich, Switzerland

Abstract

With global temperatures rising due to climate change, heat stress is becoming an increasingly prevalent condition affecting everyday navigation. Acute heat stress is known to impair complex human cognitive functioning, including memory, attention, and executive functioning. Yet, the influence of heat stress on spatial learning during navigation remains unexplored. Therefore, we conducted a VR navigation study to assess the influence of heat stress on landmark and route knowledge acquisition. Forty-six healthy adults followed a predefined route through a VR urban environment featuring intersections with and without visually salient landmarks, under either heat stress (36–38°C) or a control condition (24.5°C). We found that heat stress significantly impaired overall landmark knowledge, especially at intersections without salient landmarks, where performance failed to exceed chance level. Route knowledge, however, remained unaffected by heat stress regardless of landmark presence. These findings provide novel empirical evidence that acute heat stress selectively impairs spatial learning, highlighting the need for further research into how rising temperatures may affect human spatial cognition during navigation.

2012 ACM Subject Classification Computing methodologies → Cognitive science; Computing methodologies → Virtual reality; General and reference → Empirical studies

Keywords and phrases Heat stress, landmarks, spatial learning, VR navigation

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.29

Category Short Paper

Supplementary Material *DataPaper (related research data, materials, and analysis code)*: `https://doi.org/10.17605/osf.io/rbn4a`

1 Introduction

Imagine exploring an unfamiliar city on a sweltering summer day, as many of us do as tourists. Later that day, you tell your friend about your exciting visit by trying to recall the route you walked – which landmarks you saw along the way, and which direction you turned at each intersection. Would the heat you experienced during your walk have compromised your ability to encode and retain this spatial information? Acute heat stress is known to impair complex cognitive functions, including memory, attention, and executive functioning [8, 9, 25]. With global temperatures rising due to climate change, heat stress is becoming an increasingly prevalent environmental condition affecting not only tourists in novel environments but also everyday navigation. Yet, whether and how heat stress affects spatial learning during navigation remains unexplored.

[1] Geographic Information Visualization & Analysis (GIVA), Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

© Alessandro J. Pierro, Sara I. Fabrikant, and Armand Kapaj;
licensed under Creative Commons License CC-BY 4.0

Navigation is a complex daily activity, during which we continuously acquire spatial knowledge, broadly categorized into landmark, route, and survey knowledge [19, 23]. Landmark knowledge refers to the recognition of discrete salient features or scenes, route knowledge is knowledge about associated turning directions at landmarks or scenes, and survey knowledge integrates landmarks and routes into a common reference frame with metric information [12, 19, 23]. Landmarks serve as cognitive anchors that support the mental representation of space [5, 6, 28]. Past research has shown that wayfinders' spatial learning improves when navigating environments with visually salient landmarks compared to environments devoid of landmark information [13, 22, 27]. This benefit is attributed to modulations in attentional mechanisms whereby salient landmarks that are selectively prioritized during visual encoding are directly represented in the resulting spatial memory [10, 14, 15].

Spatial learning is also sensitive to disruption by physiological and behavioral changes induced by exposure to various stressors that can interfere with the encoding and retention of spatial information [3, 4, 21]. Acute stress has been shown to shift spatial learning from flexible, hippocampal-dependent strategies toward more rigid, habit-based stimulus-response strategies [3, 4, 21], suggesting that different components of spatial knowledge acquisition may be affected differently by stress. Acute stress can arise from psychosocial stressors, which pose a threat to the social self, such as social-evaluative or interpersonal demands, or physiological stressors of internal origin, such as illness or injury, or external origin, such as exposure to cold or hot environments [3, 25]. However, most navigation research has relied predominantly on psychosocial stressors (e.g., social-evaluative tasks) and physiological stressors such as noxious cold or physical exertion [3, 4], whereas the effects of heat stress on spatial learning remain largely unexplored. This gap is particularly relevant given rising global temperatures, especially as the effects of heat stress on cognitive functioning are well established [8, 9, 25], yet remain unknown regarding spatial learning during navigation.

Therefore, in the present study, we aimed to investigate the effect of heat stress on wayfinders' incidental spatial learning, specifically on landmark and route knowledge. Participants followed a predefined route marked by arrows that appeared on approach in a VR urban environment along 15 intersections with navigation decisions (see Figure 1). Nine of the intersections featured visually salient landmarks (e.g., churches, shops, temples), while the remaining six did not. In line with previous research on the impact of heat on cognitive performance, we hypothesized that (H1) heat stress would impair spatial learning and that (H2) the presence of salient landmarks would mitigate this negative effect on spatial learning.

2 Methods

2.1 Participants

We recruited 46 healthy adults aged 18–40 years (22 female; M age = 23.93 ± 2.36 years) with normal or corrected-to-normal vision and no history of neurological or cardiovascular disease. An a priori power analysis in G*Power for a single-predictor linear model [7], assuming a medium effect size ($f^2 = 0.20$) and a power of 0.80, estimated a minimum required sample size of $N = 42$. The sample was extended to $N = 46$ to account for potential data loss. The study was registered with the Ethics Committee of the University of Zurich (no. 25.05.14). Participants provided written informed consent and received no monetary compensation.

2.2 Design

We employed a between-subjects design with heat stress (heat vs. control) as the independent variable. Participants were evenly distributed across conditions, matched for age and gender (heat: $n = 23$, 11 female, M age = 23.91 ± 2.47 years; control: $n = 23$, 11 female, M age =

23.96 ± 2.31 years), self-reported spatial abilities (heat: $M = 3.94 \pm 0.33$; control: $M = 3.98 \pm 0.39$), and thermal sensitivity (heat: $M = 4.89 \pm 0.90$; control: $M = 4.91 \pm 1.14$), with no significant differences between conditions on any of these variables (all $ps > .770$). In each condition, participants were instructed to follow the predefined VR route and were not informed that their spatial knowledge would be tested afterward, thus ensuring incidental spatial learning. The dependent variables included performance on intersection recognition and route direction recall tasks, visual attention, electrodermal activity, and self-reported thermal comfort. In this paper, we only focus on the effect of heat stress on spatial learning (i.e., landmark and route knowledge). The remaining measures are reported in the procedure section for reproducibility purposes and will be analyzed in future work.

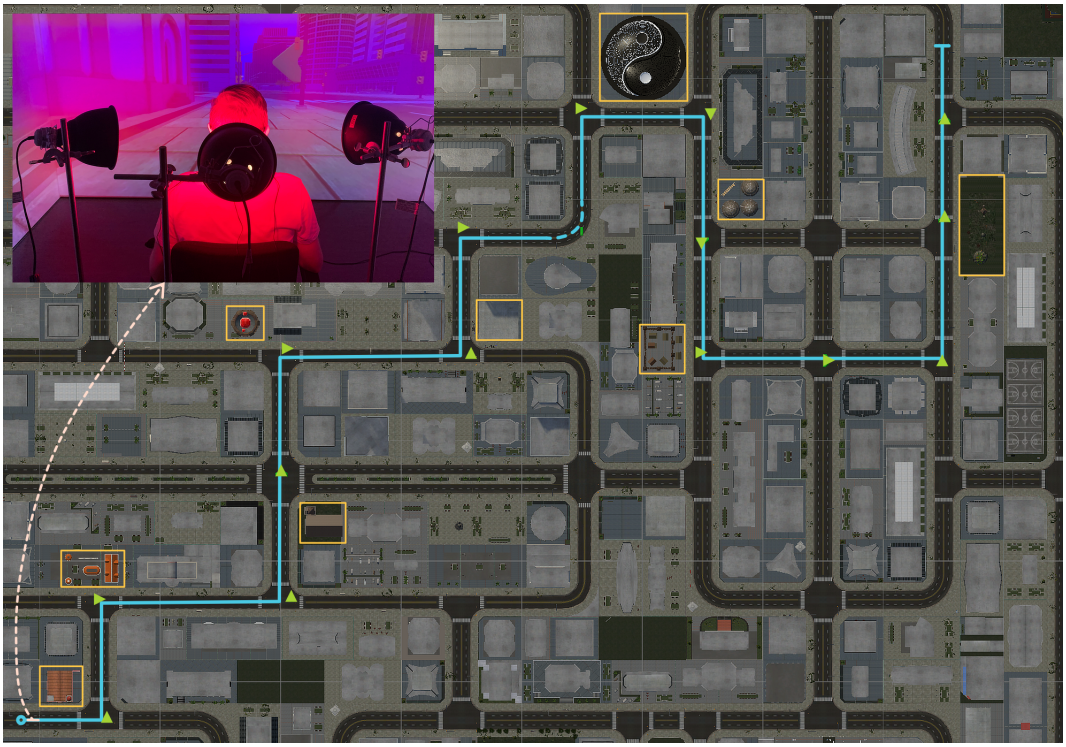


Figure 1 Overview of the VR environment depicting the predefined route (blue), the nine task-relevant landmarks (yellow), and arrows indicating the turning directions (green). The inset image shows a participant's view at an intersection enriched with a visually salient landmark.

2.3 Materials and apparatus

The VR environment depicted a city with a grid layout, modified after Bartling et al. [2], and displayed in a room-sized Cave Automatic Virtual Environment (CAVE), navigated via a foot rudder [2]. The predefined VR route comprised five left turns, five right turns, and five straight segments, for a total of 15 intersections. Nine intersections contained visually salient landmarks, such as churches, shops, and temples, whereas the remaining six had only generic high-rise buildings without distinctive visual or architectural features. To visually simulate a walk during a hot summer day, in the heat condition, we used a high midday sun illumination angle with few shaded areas, while for the control condition, the sun illumination angle was set to represent a morning walk with abundant shadows.

The heat stress was induced using three *Philips E27* infrared lamps (150 W, 230 V, 121 × 136 mm diameter) positioned to the left, right, and rear of the participant's head and neck. Each lamp was controlled by a *Ketotek KT3200PRO* thermostat, which monitored air temperature at the center of the lamp's beam. The thermostat sensors were clipped to participants' clothing, with the temperature maintained between 36°C and 42°C, to maintain consistent heat stress and to ensure safety. This resulted in an air temperature of 36–38°C around the participants' head and neck. In the control condition, the lamps remained off, and the air temperature was regulated to 24.5°C. This temperature falls within the thermal comfort ranges for temperate and air-conditioned indoor environments [1] and aligns with control temperatures used in prior heat-stress cognition research [8].

The Thermal Sensitivity Questionnaire (TSQ) was self-developed to assess participants' subjective sensitivity to heat. The TSQ measures perceived discomfort, tolerance, and behavioral and cognitive responses to heat across eight 7-point Likert scale items. Higher TSQ scores indicate higher sensitivity to heat. The TSQ showed good internal consistency in our sample, with overall Cronbach's $\alpha = .879$ and item-level values $\alpha > .852$.

The Santa Barbara Sense of Direction (SBSOD) scale was used to assess participants' spatial abilities [11]. The SBSOD measures self-reported spatial and navigational abilities, preferences, and experiences across 15 items rated on a 7-point Likert scale [11].

The intersection recognition task assessed landmark knowledge acquisition. Participants were shown images of intersections and asked if they recognized them from the traversed route. There were 15 images of intersections encountered along the route (nine with and six without landmarks), and 15 novel intersections (nine with and six without landmarks) from non-visible parts of the VR environment, for a total of 30 randomized trials.

The route direction recall task assessed route knowledge acquisition. Participants were presented with the 15 images of the traversed intersections in random order and were asked to recall the turning direction (i.e., straight, left, or right) at each intersection.

2.4 Procedure

Participants provided demographic information (age and gender) and completed the SBSOD [11] and TSQ questionnaires online, before taking part in the experiment. The responses were used to match participants across the conditions. On the experimental day, participants were welcomed in a temperature-regulated room set at 22°C where they provided written informed consent and acclimatized for 10 min. Next, they were taken to the CAVE lab and were equipped with an electrodermal activity sensor and a mobile eye-tracker. Participants were then instructed on how to use the foot pedal and practiced navigating a training VR environment for 5 min. Following this, they were instructed to walk as they would when naturally exploring a new environment and then completed the main navigation task (approx. 10–15 min). Upon task completion, the sensors were removed, and participants completed the intersection recognition task, route direction recall task, and a self-developed questionnaire on perceived thermal comfort while navigating the VR environment.

2.5 Data analyses

To assess landmark knowledge acquisition, we computed the discriminability index d' from signal detection theory (SDT; [24]) using the *psycho* package [18] in RStudio. Specifically, d' measured participants' ability to distinguish previously navigated intersections from novel

distractor intersections not encountered along the VR route [16], computed separately for intersections with (i.e., nine) and without (i.e., six) landmarks. To assess route knowledge acquisition, we used the recall accuracy of turning directions (in percent).

We employed one-tailed t-tests for each condition (heat vs. control) to assess whether participants' landmark and route knowledge – for intersections with and without landmarks – were greater than chance level, i.e., d' = 0 for landmark knowledge and 46.67% for route knowledge, accounting for intersections with two (12 intersections) or three (three intersections) turning options. A Bonferroni correction was applied to adjust the alpha level for multiple comparisons, and significance was accepted at $p < .0063$ ($\alpha/k = .05/8$ chance performance tests across tasks, conditions, and landmark presence). We also employed mixed-design ANOVAs to assess the effects of heat (heat vs. control) and landmark presence (present vs. absent) on landmark and route knowledge acquisition.

3 Results

Landmark knowledge for intersections with visually salient landmarks was significantly above chance under both heat ($t(22) = 8.47,\ p < .001,\ M_{d'} = 1.66$) and control ($t(22) = 13.82,\ p < .001,\ M_{d'} = 2.01$) conditions. For intersections without salient landmarks, performance exceeded chance in the control condition ($t(22) = 7.68,\ p < .001,\ M_{d'} = 1.04$), but did not reach the Bonferroni corrected threshold ($t(22) = 1.85,\ p = .039,\ M_{d'} = 0.26$) in the heat condition (see Figure 2a). The mixed-design ANOVA revealed a significant main effect of heat stress ($F(1, 44) = 10.50,\ p = .002,\ \eta^2_G = .131$), with participants in the control condition showing overall better landmark knowledge than those in the heat condition. We also found a significant main effect of landmark presence ($F(1, 44) = 78.21,\ p < .001,\ \eta^2_G = .396$), where intersections with salient landmarks were recognized better than those without (see Figure 2a). The interaction between heat stress and landmark presence was not significant ($F(1, 44) = 2.61,\ p = .114,\ \eta^2_G = .021$).

Route knowledge for intersections with visually salient landmarks was significantly above chance under both heat ($t(22) = 6.89,\ p < .001,\ M = 70.53\%$) and control ($t(22) = 7.37,\ p < .001,\ M = 69.08\%$) conditions. For intersections without salient landmarks, performance also exceeded chance under both heat ($t(22) = 4.72,\ p < .001,\ M = 63.04\%$) and control ($t(22) = 5.93,\ p < .001,\ M = 68.12\%$) conditions (see Figure 2b). The ANOVA revealed no main effects of heat stress ($F(1, 44) = 0.25,\ p = .619,\ \eta^2_G = .003$) and landmark presence ($F(1, 44) = 1.77,\ p = .190,\ \eta^2_G = .017$), and no significant interaction effect ($F(1, 44) = 1.06,\ p = .310,\ \eta^2_G = .010$), suggesting that neither heat stress nor landmark presence affected route direction recall (see Figure 2b).

4 Discussion

The present study examined the effect of heat stress on spatial knowledge acquisition during a navigation task in an urban VR environment. We found partial support for our hypotheses that heat stress would impair spatial learning (H1) and that the presence of visually salient landmarks in the environment would mitigate this negative effect (H2). Aligning with H1, heat stress impaired overall landmark knowledge. However, while the presence of visually salient landmarks improved overall landmark knowledge, there was no significant interaction with heat stress to support H2. Furthermore, contrary to both hypotheses, route knowledge performance was unaffected by heat stress (H1) or landmark presence (H2).

The impairment of landmark knowledge under heat stress is consistent with prior broader evidence that heat stress impairs complex cognitive functions [8, 9, 25]. Under heat stress, the additional attentional demands of thermoregulation may have reduced the cognitive resources available for encoding spatial information [25]. The resource depletion argument likely accounts also for the reduced performance in recognizing previously encountered intersections devoid of visually salient landmarks, which place greater demands on memory encoding and retrieval, as these locations offer fewer perceptual cues to anchor spatial memories [13, 27]. In contrast, intersections with visually salient landmarks may have provided sufficiently distinctive visual cues to support recognition even under reduced cognitive capacity, consistent with the role of landmarks as cognitive anchors [5, 28] that guide wayfinders' attention to task-relevant information for facilitated spatial memory formation [10, 14, 15].

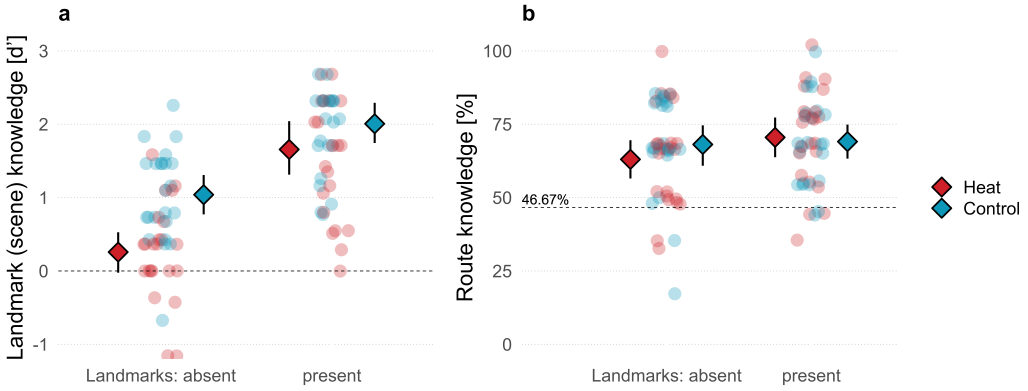


Figure 2 Heat stress impairs landmark knowledge, especially at intersections without visually salient landmarks (a), but does not affect route knowledge (b). *Note: Error bars represent 95% CIs around group means; points indicate individual data and the dashed black line shows chance level.*

The absence of heat stress effect on route knowledge suggests that the encoding of turning directions at intersections is more robust to heat stress than landmark knowledge. These findings extend previous spatial cognition research showing that route-following under psychophysiological stress relies on habit-based stimulus-response associations that place fewer demands on cognitive resources at the expense of a flexible hippocampal-dependent spatial memory system [3, 4, 21]. Notably, route knowledge acquisition was not affected by the presence or absence of visually salient landmarks. Previous studies have shown that scenes and objects are stored similarly in visual memory [17], suggesting that entire scenes function as landmarks and can be recognized even when visually distinct features are absent [20]. This may be supported by pattern completion, a memory retrieval mechanism that enables the recognition of previously encountered scenes even from partial or degraded visual cues [26]. Additionally, the selective impairment of landmark but not route knowledge is consistent with the findings that heat stress affects cognitive performance differently depending on task type and complexity, exposure duration, and air or core body temperature [8, 9, 25].

Limitations and future work

Given that we present preliminary results on the effect of heat stress on spatial learning, several limitations should be acknowledged. First, heat stress was localized to the head and neck area, which may have resulted in more moderate cognitive effects compared to

full-body hyperthermia [8, 9]. Second, we induced heat stress via infrared lamps, increasing the air temperature directly around participants to 36–38°C. Cognitive function is generally unaffected unless heat stress is sufficient to increase core body temperature beyond its homeostatic range, approximately 37°C [8, 25]. Future work should thus investigate whether increased heat intensity impacts spatial learning. Finally, the present study assessed landmark and route knowledge but not survey knowledge. Given that heat stress impairs cognitive tasks disproportionately based on their complexity [25], survey knowledge may be even more vulnerable to heat stress than landmark and route knowledge. This warrants investigation in future research. Next, we plan to analyze the collected electrodermal activity and visual attention data to assess the effect of heat stress on physiological responses, visual attention behavior, and whether visual attention patterns predict spatial learning performance [14, 15].

5 Conclusions

This study provides novel evidence that heat stress selectively impairs landmark knowledge acquisition during navigation, particularly at intersections lacking visually salient features, while leaving route knowledge intact. These findings suggest that heat stress depletes the cognitive resources required for encoding spatial features, while embodied, stimulus-response-based route knowledge remains unaffected. As temperatures continue to rise, understanding how heat stress shapes spatial learning becomes important for designing resilient urban environments and adaptable navigation assistance that supports our daily navigation needs.

References

1 F. S. Arsad, R. Hod, N. Ahmad, M. Baharom, and M. H. Ja'afar. Assessment of indoor thermal comfort temperature and related behavioural adaptations: A systematic review. *Environmental Science and Pollution Research*, 30(29):73137–73149, 2023. doi:10.1007/s11356-023-27089-9.

2 M. Bartling, A. Kapaj, B. Cheng, Z. Xu, A. L. Griffin, and S. I. Fabrikant. Unlocking context-aware mobile map adaptation: A VR navigation user study. *Cartography and Geographic Information Science*, 53(1):56–74, 2026. doi:10.1080/15230406.2025.2512075.

3 A. P. Boone, T. Bullock, M. H. MacLean, T. Santander, J. Raymer, A. Stuber, L. Jimmons, G. N. Okafor, S. T. Grafton, M. B. Miller, B. Giesbrecht, and M. Hegarty. Resilience of navigation strategy and efficiency to the impact of acute stress. *Spatial Cognition & Computation*, 24(3):195–226, 2024. doi:10.1080/13875868.2024.2321920.

4 T. I. Brown, S. A. Gagnon, and A. D. Wagner. Stress disrupts human hippocampal-prefrontal function during prospective spatial navigation and hinders flexible behavior. *Current Biology*, 30(10):1821–1833.e8, 2020. doi:10.1016/j.cub.2020.03.006.

5 E. Chan, O. Baumann, M. Bellgrove, and J. Mattingley. From objects to landmarks: The function of visual location information in spatial navigation. *Frontiers in Psychology*, 3, 2012. doi:10.3389/fpsyg.2012.00304.

6 R. A. Epstein and L. K. Vass. Neural systems for landmark-based wayfinding in humans. *Philos Trans R Soc Lond B Biol Sci*, 369(1635):20120533, 2014. doi:10.1098/rstb.2012.0533.

7 F. Faul, E. Erdfelder, AG. Lang, and A. Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007. doi:10.3758/BF03193146.

8 N. Gaoua, J. Grantham, S. Racinais, and F. El Massioui. Sensory displeasure reduces complex cognitive performance in the heat. *Journal of Environmental Psychology*, 32(2):158–163, 2012. doi:10.1016/j.jenvp.2012.01.002.

9 N. Gaoua, S. Racinais, J. Grantham, and F. El Massioui. Alterations in cognitive performance during passive hyperthermia are task dependent. *International Journal of Hyperthermia*, 27(1):1–9, 2011. doi:10.3109/02656736.2010.516305.

10 S.N. Hamid, B. Stankiewicz, and M. Hayhoe. Gaze patterns in navigation: Encoding information in large-scale environments. *Journal of Vision*, 10(12):28, 2010. doi:10.1167/10.12.28.

11 M. Hegarty, A. E. Richardson, D. R. Montello, K. Lovelace, and I. Subbiah. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002. doi:10.1016/S0160-2896(02)00116-2.

12 T. Ishikawa and D. R. Montello. Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, 52(2):93–129, 2006. doi:10.1016/j.cogpsych.2005.08.003.

13 P. Jansen-Osmann and P. Fuchs. Wayfinding behavior and spatial knowledge of adults and children in a virtual environment. *Experimental Psychology*, 53(3):171–181, 2006. doi:10.1027/1618-3169.53.3.171.

14 A. Kapaj, C. Hilton, S. Lanini-Maggi, and S. I. Fabrikant. The influence of landmark visualization style on task performance, visual attention, and spatial learning in a real-world navigation task. *Spatial Cognition & Computation*, 24(4):227–267, 2024. doi:10.1080/13875868.2024.2328099.

15 A. Kapaj, S. Lanini-Maggi, C. Hilton, B. Cheng, and S. I. Fabrikant. How does the design of landmarks on a mobile map influence wayfinding experts' spatial learning during a real-world navigation task? *Cartography and Geographic Information Science*, 50(2):197–213, 2023. doi:10.1080/15230406.2023.2183525.

16 A. Kapaj, E. Lin, and S. Lanini-Maggi. The effect of abstract vs. realistic 3D visualization on landmark and route knowledge acquisition (Short Paper). In *15th International Conference on Spatial Information Theory (COSIT 2022)*, volume 240, pages 15:1–15:8, 2022. doi:10.4230/LIPIcs.COSIT.2022.15.

17 T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, 21(11):1551–1556, 2010. doi:10.1177/0956797610385359.

18 D. Makowski. The psycho package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, 3(22):470, 2018. doi:10.21105/joss.00470.

19 D. R. Montello. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In *Spatial and temporal reasoning in geographic information systems*, pages 143–154. Oxford University Press, 1998.

20 D. R. Montello. Landmarks are Exaggerated. *KI - Künstliche Intelligenz*, 31(2):193–197, 2017. doi:10.1007/s13218-016-0473-5.

21 L. Schwabe, M. S. Oitzl, C. Philippsen, S. Richter, A. Bohringer, W. Wippich, and H. Schachinger. Stress modulates the use of spatial versus stimulus-response learning strategies in humans. *Learning & Memory*, 14(1-2):109–116, 2007. doi:10.1101/lm.435807.

22 G. Sharma, Y. Kaushal, S. Chandra, V. Singh, A. P. Mittal, and V. Dutt. Influence of landmarks on wayfinding and brain connectivity in immersive virtual reality environment. *Frontiers in Psychology*, 8, 2017. doi:10.3389/fpsyg.2017.01220.

23 A. W. Siegel and S. H. White. The development of spatial representations of large-scale environments. In *Advances in Child Development and Behavior*, volume 10, pages 9–55. JAI, 1975. doi:10.1016/S0065-2407(08)60007-5.

24 H. Stanislaw and N. Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31:137–149, 1999. doi:10.3758/BF03207704.

25 L. Taylor, S. L. Watkins, H. Marshall, B. J. Dascombe, and J. Foster. The impact of different environmental conditions on cognitive function: A focused review. *Frontiers in Physiology*, 6, 2016. doi:10.3389/fphys.2015.00372.

26 P. Vieweg, M. Stangl, L. R. Howard, and T. Wolbers. Changes in pattern completion – a key mechanism to explain age-related recognition memory deficits? *Cortex*, 64:343–351, 2015. doi:10.1016/j.cortex.2014.12.007.

27 D. Waller and Y. Lippa. Landmarks as beacons and associative cues: Their role in route learning. *Memory & Cognition*, 35(5):910–924, 2007. doi:10.3758/BF03193465.

28 D. Yesiltepe, R. Conroy Dalton, and A. Ozbil Torun. Landmarks in wayfinding: A review of the existing literature. *Cognitive Processing*, 22(3):369–410, 2021. doi:10.1007/s10339-021-01012-x.

From Feature Attribution to Spatial Information Structure: Interpreting Urban Pedestrian Flows with Explainable AI

Atakilti Kiros
Faculty of Engineering, Ariel University, Israel

Achituv Cohen
Faculty of Engineering, Ariel University, Israel

Yuri Ribakov
Faculty of Engineering, Ariel University, Israel

Israel Klein
Ariel University, Israel

Abstract

Machine-learning models can predict pedestrian flows with high accuracy, but they provide limited insight into how urban environments organize movement. This paper introduces an interpretive framework that links explainable artificial intelligence with Spatial Information Theory by conceptualizing feature attribution as evidence of spatial information structure – the relative salience of temporal, morphological, accessibility, and environmental dimensions in shaping pedestrian movement. Using pedestrian sensor data from Zurich and Berlin, we train comparable CatBoost models and analyze SHAP-based feature contributions. We examine whether different urban contexts encode movement in systematically distinct ways. The results reveal clear contrasts: Berlin is dominated by temporal dynamics and situational conditions, whereas Zurich shows stronger influence from urban morphology and accessibility. We interpret these differences as model-based evidence that cities organize movement-related information differently and demonstrate that a compact SHAP-based comparative design provides a practical methodological bridge between GeoAI and the study of spatial representation.

2012 ACM Subject Classification Computing methodologies → Machine learning; Information systems → Geographic information systems; Applied computing → Environmental sciences; Information systems → Spatial-temporal systems; Human-centered computing → Visualization techniques

Keywords and phrases GeoAI, explainable artificial intelligence, SHAP, pedestrian mobility, spatial information theory, urban analytics, feature attribution, spatial data science

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.30

Category Short Paper

1 Introduction

Understanding how urban space structures human movement remains a central challenge in geography, urban science, and Spatial Information Theory (SIT). While machine-learning models can accurately predict pedestrian flows using spatial, temporal, and environmental data, they offer limited insight into how movement is organized within urban environments. Recent advances in GeoAI demonstrate strong predictive capabilities for modeling complex urban dynamics, particularly through the integration of spatial, temporal, and multimodal data using machine learning and deep learning approaches [5, 7, 2]. Explainable artificial intelligence (XAI) provides a pathway to address this limitation by enabling interpretation of model behavior in geospatial contexts. Recent work has also explored explainable spatio-temporal models in urban contexts, highlighting the importance of interpretability for

© Atakilti Kiros, Achituv Cohen, Yuri Ribakov, and Israel Klein;
licensed under Creative Commons License CC-BY 4.0

understanding complex urban systems [1]. SHapley Additive exPlanations (SHAP) enables the quantification of feature contributions to model predictions [8], offering a means to examine how different dimensions of urban space relate to movement. However, the potential of XAI to contribute to SIT remains underexplored. From a SIT perspective, urban space can be understood as an information system that encodes relationships such as accessibility, morphology, and environmental context, which guide human behavior [4, 9, 6]. Despite this conceptual foundation, empirical methods for approximating such representations from behavioral data remain limited.

This paper addresses this gap by introducing the concept of spatial information structure, defined as the relative configuration of temporal, morphological, accessibility, and environmental dimensions that organize pedestrian movement. Rather than treating variables as independent predictors, we interpret them as components of a relational system through which urban space constrains and affords activity. We operate this concept using SHAP-based analysis of pedestrian flow data from two European cities, Zurich and Berlin. By comparing feature contributions across cities under a consistent modeling framework, we examine whether different urban contexts exhibit distinct configurations of spatial information.

2 Conceptual Framework: Spatial Information Structure

SIT conceptualizes urban environments as information-rich systems in which multiple dimensions such as built form, activity patterns, environmental conditions, and temporal rhythms jointly shape movement [3, 9, 6]. Building on this perspective, we use the term spatial information structure to refer to the relational organization of temporal, morphological, accessibility, and environmental dimensions within urban space. In this view, these dimensions are not treated as isolated predictors, but as interconnected components of a spatial system that may constrain and afford movement. Observed pedestrian flows are therefore considered not as the structure itself, but as a behavioral expression through which aspects of this underlying spatial information structure may be empirically approximated.

In the empirical implementation, features are grouped into a small set of conceptually distinct dimensions: temporal dynamics, urban morphology, accessibility, and environmental conditions. Temporal features capture cyclical patterns of activity; morphological features represent structural constraints of the built environment; accessibility features encode proximity and connectivity to services and transport; and environmental variables describe climatic and situational influences. These dimensions provide an operational approximation of how different types of urban information contribute to movement, consistent with data-driven approaches in urban analytics and GeoAI [3, 10]. Because the spatial information structure is not directly observable, it must be inferred indirectly from the relationship between environmental descriptors and observed behavior.

To support this inference, we use SHAP values to quantify the contribution of individual features to the predictions of the model [8]. Importantly, SHAP values are not interpreted as direct measurements of spatial structure but as model-based evidence that can be aggregated to approximate the relative salience of different information dimensions. This approach aligns with recent work in explainable GeoAI, which emphasizes interpreting model behavior to better understand spatial processes [1], rather than focusing solely on predictive performance in data-driven urban models [11].

This shifts the focus from predictive accuracy to identifying which dimensions of spatial information organize movement. Spatial information structures may vary across cities, requiring consistent models and features for comparison. It is defined as the relative importance of temporal, morphological, accessibility, and environmental contributions.

3 Study Design and Methods

The analysis is based on pedestrian sensor data from Zurich and Berlin collected throughout the full year of 2024. The dataset comprises approximately 60 sensors in Zurich and 33 sensors in Berlin, each providing hourly pedestrian counts across diverse urban contexts. Sensor data were obtained from municipal open data platforms. Environmental variables were derived from meteorological sources such as Meteostat, while urban morphology, accessibility, and network features were computed using OpenStreetMap (OSM) and remote sensing products such as Sentinel-2.

To ensure cross-city comparability, a harmonized set of 21 features was constructed using identical preprocessing pipelines. These features capture four conceptual dimensions: temporal dynamics, urban morphology, accessibility, and environmental conditions. Spatial variables were computed using standardized buffer-based and network-based methods, including density estimation, proximity measures, and centrality metrics. Temporal dynamics are represented as two cyclic features, Hour (cos) $= \cos(2\pi \cdot \text{hour}/24)$ and Hour (sin) $= \sin(2\pi \cdot \text{hour}/24)$, which preserve cyclic proximity between adjacent hours and allow the model to learn diurnal phase without a discontinuity at midnight. A detailed feature-level description, including data sources and operationalisation, is provided in Appendix A. To facilitate meaningful comparison, identical preprocessing procedures, feature definitions, and model settings were applied across both cities. This controlled design enables differences in feature contributions to be interpreted as differences in the organization of movement-relevant urban information rather than artifacts of model specification.

Pedestrian flows were modelled using CatBoost, with separate models trained for each city under identical configurations. We intentionally excluded autoregressive lag features to reduce simple temporal persistence effects that could dominate feature attribution and obscure relationships between movement and static urban context; this choice prioritises the interpretability of spatial and contextual predictors over maximal short-term forecasting skill. The modelling workflow consists of four stages: data integration and preprocessing; multimodal feature engineering; model training and validation with standardised splits; and post hoc interpretation using SHAP values.

Robustness was assessed across five random seeds and five independent 80/20 train–validation splits, with feature importance rankings aggregated to assess their stability across model runs.

4 Results

4.1 Model Performance

The predictive models achieve strong performance in both study areas, indicating that the selected feature set captures meaningful and spatially relevant patterns in pedestrian mobility.

As shown in Figure 1, the observed-vs.-predicted scatterplots demonstrate that the models explain a substantial proportion of the variance in pedestrian flows, with $R^2 = 0.76$ in Zurich and $R^2 = 0.80$ in Berlin. In both cities, the points cluster closely around the 1:1 reference line, showing that the models track the overall magnitude and distribution of pedestrian counts well, while still allowing for some dispersion at higher flow values. This pattern suggests that the combination of temporal, environmental, and urban form variables provides a robust basis for modeling pedestrian dynamics across distinct urban contexts.

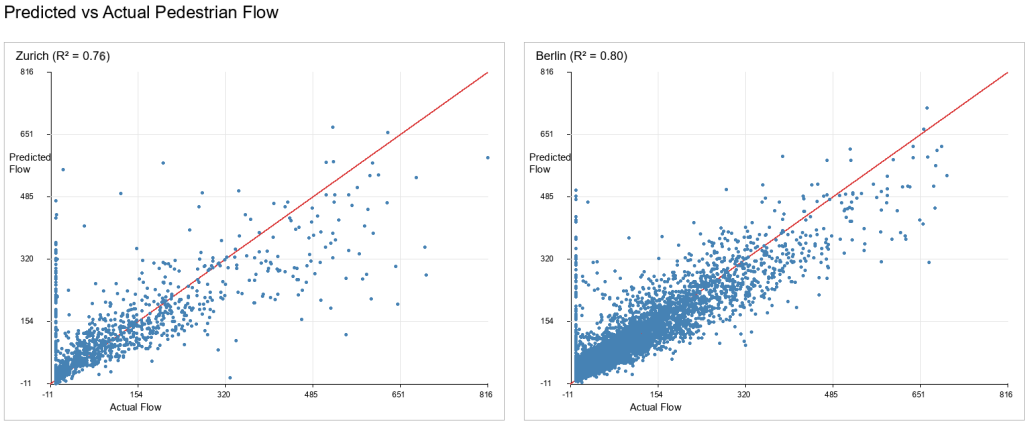


Figure 1 Predicted versus observed pedestrian flows in Zurich and Berlin, showing strong model performance ($R^2 = 0.76$ and 0.80).

4.2 Global Feature Contributions

Across both cities, temporal variables dominate the model, with cyclical hour-of-day features exhibiting the highest contributions. In both Zurich and Berlin, hour (cos) and hour (sin) are the most influential predictors, indicating that pedestrian flows are primarily structured by consistent diurnal rhythms. Beyond these shared temporal effects, non-temporal contributions differ systematically across the two urban contexts.

Table 1 Top five feature contributions based on mean absolute SHAP values and normalized importance.

Zurich			*Berlin*		
Feature	SHAP	Normalized	Feature	SHAP	Normalized
Hour (cos)	11.36	1.00	Hour (cos)	39.77	1.00
Facade continuity	9.14	0.81	Hour (sin)	27.13	0.68
Hour (sin)	8.67	0.76	Temperature	25.57	0.64
Built volume density	8.06	0.71	Street enclosure	23.68	0.60
Transit stop distance	6.65	0.59	Weekend	15.81	0.40

As shown in Table 1, Zurich is characterized by stronger morphological and accessibility-related contributions, particularly facade continuity (0.81), built volume density (0.71), and transit stop distance (0.59). In contrast, Berlin exhibits greater influence from temperature (0.64) and street enclosure (0.60), while accessibility plays a comparatively smaller role.

These patterns indicate that, although temporal dynamics provide a common structural basis, the relative salience of non-temporal features varies across cities. From a Spatial Information Theory perspective, this suggests that pedestrian movement reflects city-specific spatial information structures, with Zurich more strongly shaped by built form and accessibility, and Berlin by environmental and enclosure conditions.

4.3 Local Contribution Patterns

The SHAP beeswarm plots figure 2 provide a more detailed view of how influential features contribute to pedestrian flow predictions in each city. As shown in the global importance analysis, the temporal variables have the strongest impact in both Zurich and Berlin.

This pattern is particularly clear in Berlin, where higher values of Hour (cos) are generally associated with negative SHAP values and lower values with positive SHAP values, reflecting the cyclical nature of daily pedestrian rhythms.

Figure 2 SHAP beeswarm plots for Zurich and Berlin illustrating the distribution and direction of feature effects across observations.

In Zurich, non-temporal variables such as facade continuity, built volume density, temperature, and transit stop distance also show substantial contribution ranges, indicating that they play an important role in the model. However, their effects are less clearly directional, suggesting more context-dependent relationships. In Berlin, temperature shows a much clearer pattern, with lower temperatures contributing negatively and higher temperatures contributing positively to the model output. Built volume density also appears to have a clearer directional effect, while facade continuity and transit stop distance play a comparatively smaller role than in Zurich.

These results suggest that although temporal rhythms dominate in both cities, the structure of non-temporal influences differ meaningfully between them.

4.4 Comparative Structures

Although the feature-level analysis in Section 4.2 highlights differences in the importance of individual predictors, the normalized contributions shown in Figure 3 reveal higher-level structural patterns. Temporal variables dominate in both cities, yet the relative importance of non-temporal features differs systematically.

Zurich exhibits a more distributed configuration of influence across morphological and accessibility dimensions, whereas Berlin shows a more concentrated reliance on temporal and environmental signals. This suggests that pedestrian movement in Zurich is shaped by a more heterogeneous spatial information structure, while in Berlin it is governed by a narrower set of dominant factors. From a SIT perspective, these patterns indicate that urban environments may encode behaviorally relevant information in fundamentally different ways across cities.

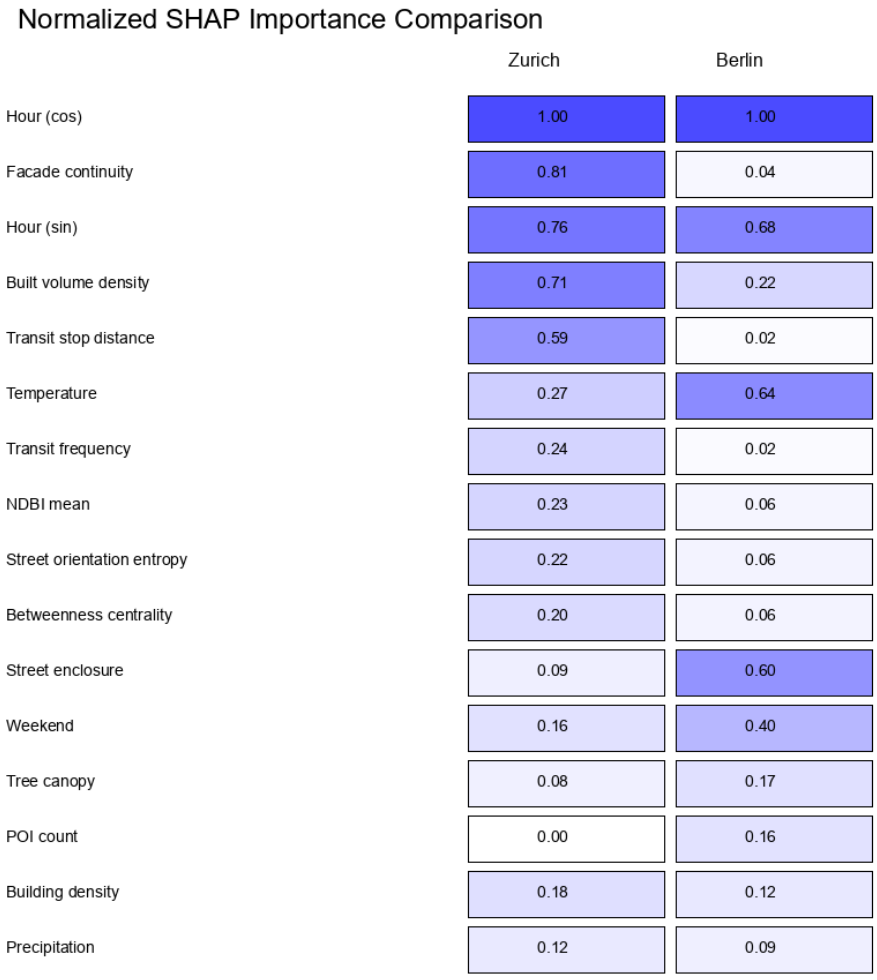


Figure 3 Normalized SHAP importance profiles for Zurich and Berlin.

4.5 Robustness of Feature Contribution Patterns

Across both cities, temporal variables dominate the model, with cyclical hour-of-day features exhibiting the highest contributions. In both Zurich and Berlin, hour (cos) and hour (sin) are the most influential predictors, indicating that pedestrian flows are primarily structured by consistent diurnal rhythms. Beyond these shared temporal effects, non-temporal contributions differ systematically across the two urban contexts.

These patterns indicate that, while temporal dynamics provide a common baseline, the importance of non-temporal features varies across cities. From a SIT perspective, this suggests that pedestrian movement reflects city-specific spatial information structures, with Zurich shaped more by built form and accessibility, and Berlin by environmental and enclosure conditions.

5 Discussion and Conclusion

The results indicate that while temporal dynamics provide a common structural basis for pedestrian flows in both cities, the relative importance of non-temporal features varies in ways that reflect distinct urban configurations. Zurich exhibits a more distributed contribution structure, with stronger influence from morphological and accessibility-related variables such

Table 2 Robustness of normalized SHAP feature contributions across repeated model runs.

	Zurich					*Berlin*			
Feature	Mean	SD	Rank	Range	Feature	Mean	SD	Rank	Range
Hour (cos)	1.00	0.00	1.0	1–1	Hour (cos)	1.00	0.00	1.0	1–1
Facade continuity	0.81	0.04	2.2	2–3	Hour (sin)	0.68	0.03	2.1	2–3
Hour (sin)	0.76	0.03	2.8	2–4	Temperature	0.64	0.05	3.0	2–4
Built volume density	0.71	0.05	4.0	3–5	Street enclosure	0.60	0.04	4.0	3–5
Transit stop distance	0.59	0.06	5.1	4–6	Weekend	0.40	0.05	5.2	4–6

as facade continuity, built volume density, and transit stop distance. In contrast, Berlin shows a more concentrated pattern in which environmental and enclosure-related factors, particularly temperature and street enclosure, are more prominent. These differences suggest that, beyond shared temporal rhythms, cities encode movement-relevant information through distinct configurations of urban form, accessibility, and environmental context.

These findings should not be interpreted as evidence of direct causal relationships, but rather as model-based approximations derived from SHAP. From a Spatial Information Theory perspective, their significance lies in revealing how different dimensions of urban space become salient within the model's learned representation of movement. Rather than identifying independent determinants of pedestrian activity, the analysis highlights how multiple dimensions jointly organize behaviorally meaningful information. In this sense, the observed contribution patterns can be understood as an empirical approximation of a city's spatial information structure. This interpretation remains subject to limitations. The inferred structures are derived from observational data and may reflect confounding factors, feature dependencies, or sensor distribution effects. The results should therefore be regarded as exploratory rather than definitive. Nevertheless, they provide a theoretically grounded and empirically interpretable basis for comparing how different urban environments organize movement-related information.

In summary, this paper introduced spatial information structure as a framework linking explainable GeoAI with Spatial Information Theory. The comparative analysis of Zurich and Berlin demonstrates that, while temporal dynamics form a shared foundation, non-temporal contributions vary systematically across cities, reflecting distinct informational configurations. By interpreting feature contributions as components of a relational system, the study moves beyond prediction toward a more interpretable understanding of how urban environments encode behaviorally relevant information. The comparison should also be interpreted in light of substantial differences in city size, population, and urban structure between Zurich and Berlin. Rather than comparing equivalent urban systems, the analysis examines whether a common modeling framework can reveal distinct configurations of movement-relevant information across contrasting urban contexts. A key direction for future work is to analyze the spatio-temporal structure of prediction errors to better understand the limits of the inferred spatial information structure. Identifying where and when the model fails may reveal missing contextual or event-driven factors and support the integration of context-aware variables, improving interpretability and generalization.

References

1 A. Alamelh. Explain-city: Explainable spatio-temporal deep learning for urban resilience and risk governance. *IEEE Access*, 13:181989–182004, 2025. doi:10.1109/ACCESS.2025.3623066.

2 S. Avram. Advancing urban analytics: Geoai applications in spatial decision-making and sustainable cities. *Urban Science*, 10, 2026. doi:10.3390/urbansci10030148.

3 Michael Batty. *The New Science of Cities*. MIT Press, 2013.

4 Max J. Egenhofer and David M. Mark. Naive geography. In Andrew U. Frank and Werner Kuhn, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, pages 1–15. Springer, Berlin, 1995. doi:10.1007/3-540-60392-1_1.

5 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34:625–636, 2020. doi:10.1080/13658816.2019.1684500.

6 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26:2267–2276, 2012. doi:10.1080/13658816.2012.722637.

7 Peng Liu, Yu Zhang, and Filip Biljecki. Explainable spatially explicit geospatial artificial intelligence in urban analytics. *Environment and Planning B: Urban Analytics and City Science*, 51:1104–1123, 2024. doi:10.1177/23998083231204689.

8 Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

9 Daniel R. Montello. Cognitive map-design research in the twentieth century: Theoretical and empirical approaches. *Cartography and Geographic Information Science*, 31(4):283–304, 2004.

10 Fivos Papadimitriou. *Spatial Artificial Intelligence*. SpringerBriefs in Applied Sciences and Technology. Springer, Cham, 2025. doi:10.1007/978-3-031-82136-3.

11 Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A Appendix A: Feature Definition and Data Sources

Table 3 Feature taxonomy, data sources, and operationalisation of the 21 predictors.

Category	Feature	Data Source	Operationalization
Temporal Dynamics			
Temporal	hour_cos	Sensor timestamp	Cosine encoding of hour-of-day
Temporal	hour_sin	Sensor timestamp	Sine encoding of hour-of-day
Temporal	is_weekend	Calendar data	Binary weekend indicator
Environmental / Weather			
Environment	temp	Meteostat / NOAA	Hourly air temperature
Environment	precip	Meteostat / NOAA	Hourly precipitation intensity
Urban Morphology (Built Form)			
Built form	built_volume_density_200m	OpenStreetMap	Built volume within 200 m buffer
Built form	building_density_200m	OpenStreetMap	Building footprint density (200 m)
Built form	facade_continuity_proxy	OpenStreetMap	Proxy of street wall continuity
Built form	street_enclosure_index_proxy	OpenStreetMap	Enclosure ratio (height/width)
Built form	street_canyon_ratio_mean_200m	OpenStreetMap	Mean height-to-width ratio (200 m)
Urban Green / Environmental Context			
Green context	sensor_canopy_pct	Remote sensing	Tree canopy percentage
Green context	sensor_ndvi_mean	Sentinel-2	Mean NDVI (buffer)
Green context	ndbi_mean_s2	Sentinel-2	Mean built-up index
Accessibility / Transport			
Accessibility	transit_stop_distance	OpenStreetMap	Distance to nearest transit stop
Accessibility	transit_frequency_index	GTFS / OSM	Frequency-weighted accessibility index
Accessibility	bus_stop_density_400m	OpenStreetMap	Bus stop density (400 m)
Urban Activity / Land Use			
Land use	poi_count_300m	OpenStreetMap	POI count within 300 m
Land use	access_to_services_index	OpenStreetMap	Composite service accessibility index
Street Network Structure			
Network	intersection_density_300m	OpenStreetMap	Intersection density (300 m)
Network	betweenness_centrality	OpenStreetMap	Network centrality measure
Network	street_orientation_entropy	OpenStreetMap	Orientation diversity entropy

Spatiotemporal Adaptive Learning for Flow-Based Graph Neural Networks

Fengjiao Li[1]
School of Geographical and Earth Sciences, University of Glasgow, UK

Meiliu Wu
School of Geographical and Earth Sciences, University of Glasgow, UK

Ana Basiri
School of Geographical and Earth Sciences, University of Glasgow, UK

Abstract

Spatial systems can often be represented in more than one relational way. For the same urban system, geographic adjacency, road connectivity, and mobility interaction provide different relational representations of how places are connected. However, most graph-based models either assume a single graph or combine multiple graphs without distinguishing their relative importance. This paper treats multi-graph spatiotemporal modelling as a problem of spatial perspective selection. We introduce adaptive spatial weights (α) to learn the relative importance of predefined relational graphs, and temporal weights (β) in the loss function to distribute emphasis across forecasting horizons. Using COVID-19 dynamics in Glasgow and Edinburgh as a case study, we find that learned spatial weighting patterns differ across cities. Glasgow shows a more balanced configuration, whereas Edinburgh places stronger emphasis on road connectivity. Compared with single-graph and equal-weight baselines, the adaptive model achieves comparable performance with improved stability while making graph weighting explicit. These results suggest that multi-graph weighting is a context-dependent process of selecting among alternative relational representations.

2012 ACM Subject Classification Theory of computation → Network flows; Computing methodologies → Temporal reasoning; Computing methodologies → Spatial and physical reasoning; Computing methodologies → Artificial intelligence; Information systems → Geographic information systems

Keywords and phrases spatiotemporal adaptive learning, graph neural networks, multi-graph representation, GeoAI, flow-based networks

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.31

Category Short Paper

Funding This research was supported by the China Scholarship Council (CSC).

Acknowledgements The authors also acknowledge the use of publicly available data provided by Public Health Scotland (COVID-19 in Scotland dataset) and the Urban Big Data Centre (UBDC) at the University of Glasgow (mobility dataset).

Declaration on GenAI/LLM use ChatGPT was used to assist in the creation of Figure 1. The figure was generated for illustrative purposes, while the underlying scientific ideas, model design, and interpretation were developed by the authors. The generated visual content was reviewed, edited, and approved by the authors to ensure that it accurately reflects the content of the paper.

1 Introduction

Spatial systems are inherently relational and can be represented in multiple ways. The meaning of a location is shaped not only by its attributes, but also by the relations through which it is defined. In spatial information theory, space is increasingly viewed as a relational

[1] Corresponding author.

© Fengjiao Li, Meiliu Wu, and Ana Basiri;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 31; pp. 31:1–31:9
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

construct rather than a fixed geometric container [2]. Similarly, research in spatial cognition shows that spatial representation depends on perspective, scale, and task [10]. As a result, different relational descriptions, such as geographic adjacency, infrastructural connectivity, and behavioural interaction, can provide valid but distinct views of the same system. This idea is also reflected in multilayer and multiplex network research, which formalises how multiple relational layers may jointly shape system dynamics [8, 3, 4]. In GIScience, related work has likewise emphasised that spatial representation is context-dependent rather than uniquely determined, and that different abstractions of the same system may coexist for different purposes. Together, these perspectives suggest that spatial systems admit multiple meaningful relational encodings rather than a single privileged structure.

Graph Neural Networks (GNNs) have become central tools for modelling relational structure in spatial and spatiotemporal systems, including applications such as traffic forecasting and epidemic modelling [9, 18, 16]. Spatiotemporal models such as STGCN and DCRNN have demonstrated strong forecasting performance [18, 9], while subsequent research has introduced attention mechanisms, graph learning, and multi-graph extensions to improve predictive flexibility [6, 17, 5, 15]. Most existing approaches either assume a single graph or combine multiple graphs using implicit mechanisms such as attention or graph learning [5, 15, 6]. However, these approaches primarily treat graph construction and graph combination as modelling or optimisation choices. Even when multiple relational graphs are used, fusion is usually implemented as weighted aggregation or attention-based combination, rather than being interpreted as explicit selection among coexisting spatial perspectives.

This issue is particularly important in real spatial systems, where different graphs encode different spatial logics. Geographic adjacency reflects spatial proximity, road networks capture infrastructural constraints, and mobility flows represent behavioural interaction across areas [1, 12]. These representations are not interchangeable, but instead provide complementary perspectives on the same system. From this perspective, multi-graph modelling can be understood not only as a problem of technical fusion, but also as a problem of representational selection. While previous studies have largely focused on improving predictive accuracy, comparatively less attention has been paid to understanding how alternative relational representations should be prioritised and what such prioritisation may reveal about the spatial processes being modelled.

Our work builds on these two lines of research. First, we recognise that multiple relational encodings of the same spatial system can be theoretically meaningful, with each graph representing a distinct spatial perspective and a corresponding conceptual flow mechanism. Second, we adopt a spatiotemporal graph learning framework in which these perspectives can be combined adaptively. Unlike most previous studies, which primarily view multi-graph learning as a problem of graph fusion for improving predictive accuracy, we interpret graph weighting as a process of spatial perspective selection. In this view, learned graph weights provide not only a mechanism for prediction, but also an interpretable indication of which relational representations may be more informative in a given context. Predictive performance therefore serves as supporting evidence for a broader representational question concerning the relative importance of alternative spatial perspectives.

In this paper, we frame multi-graph spatiotemporal modelling as a problem of *spatial perspective selection*. We make explicit how different relational perspectives are weighted within a unified model. To this end, we introduce an adaptive spatial weighting mechanism α to learn the relative contribution of different graphs. We also incorporate temporal weighting β to account for differences across prediction horizons. Together, these provide a spatiotemporal weighting formulation, although our primary focus remains on the role of

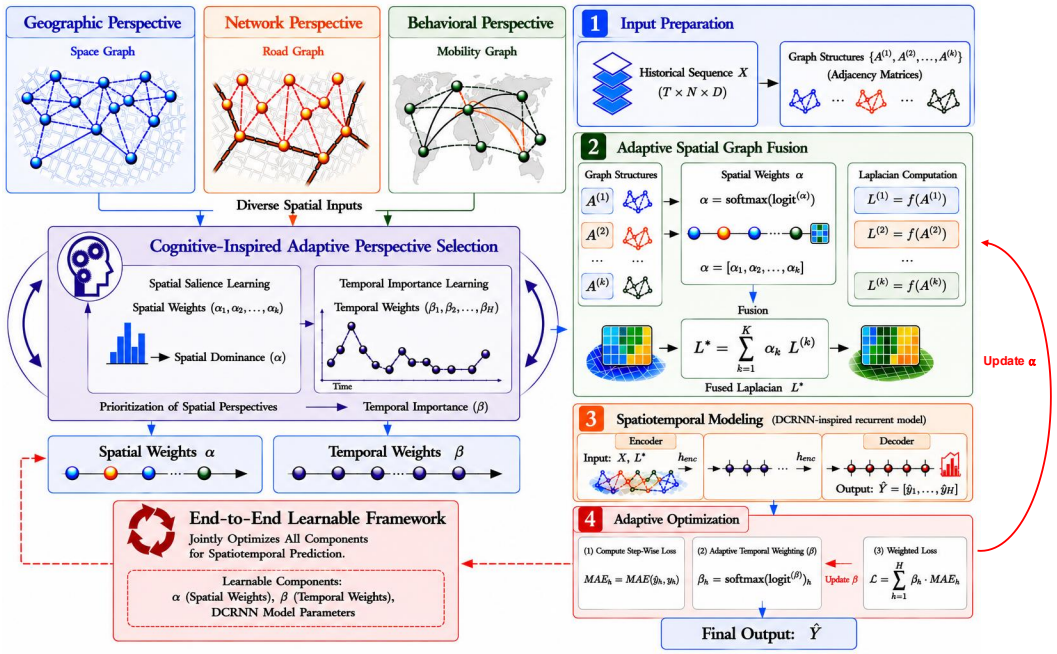


Figure 1 Overview of the proposed spatiotemporal adaptive learning framework.

spatial weighting in representing different perspectives. To illustrate this idea, we implement the framework within a spatiotemporal graph neural model and evaluate it using epidemic dynamics in Glasgow and Edinburgh. Rather than focusing solely on predictive performance, we examine how learned spatial weights vary across contexts, providing insight into how different relational representations contribute in different urban settings.

This paper makes three contributions. First, it frames multi-graph modelling as a problem of spatial perspective selection. Second, it provides a practical formulation using α and β to make relational weighting explicit. Third, it reveals empirical findings that spatial weighting patterns differ across cities, suggesting that relational importance is context-dependent and potentially linked to different underlying flow mechanisms represented by each graph.

2 A Novel Approach: Spatiotemporal Adaptive Learning

Existing spatiotemporal graph models often combine spatial fusion and temporal aggregation within a single adaptive process, making it difficult to distinguish weighting across relational graphs from weighting across prediction horizons. Here, we separate these two forms of weighting within a unified framework (Figure 1). Multiple spatial graphs are treated as distinct relational perspectives, including geographic adjacency, infrastructure connectivity, and mobility interaction. Spatial weighting is defined through α, which determines the contribution of each graph and satisfies $\sum_{k=1}^{K} \alpha_k = 1$, producing a fused spatial operator L^*. Within the proposed framework, these weights can be interpreted as model-based proxies for the relative importance of different spatial perspectives.

Temporal weighting is introduced through β, which distributes emphasis across prediction horizons. While α governs spatial combination, β controls the weighting of prediction errors over time. Taken together, they define a spatiotemporal weighting formulation that separates spatial perspective prioritisation from temporal focus allocation.

Table 1 Conceptual correspondence between relational graphs and spatial perspectives.

Graph	Spatial Perspective	Flow Mechanism
Geographic adjacency	Metric continuity	Local spillover
Road connectivity	Infrastructural constraint	Transport-mediated diffusion
Mobility interaction	Behavioural interaction	Movement-driven transmission

Conceptually, the proposed framework distinguishes two related but different questions. The first concerns which spatial perspectives should receive greater emphasis when multiple relational representations coexist. The second concerns which forecasting horizons should receive greater emphasis during model optimisation. By separating adaptive spatial weighting (α) from adaptive temporal weighting (β), the framework allows learned graph weights to be interpreted as model-based indicators of relative spatial importance rather than as components of a generic attention mechanism. This distinction helps clarify how spatial and temporal information contribute to the overall learning process.

3 Data and Methods

3.1 Data

The empirical analysis focuses on Glasgow and Edinburgh. Daily COVID-19 confirmed case counts at the InterZone level (March 2020–February 2023) were obtained from Public Health Scotland [13]. Origin–destination mobility matrices were obtained from the Urban Big Data Centre (UBDC) [14], and road and rail networks were downloaded from OpenStreetMap via Geofabrik [11]. These datasets support the construction of geographic, infrastructural, and mobility-based graphs.

Let $X_t \in \mathbb{R}^N$ denote infection rates at time t, and $Y \in \mathbb{R}^{H\times N}$ denote future values over H horizons. A history window $T = 7$ is used to predict $H = 7$ days. Inputs are standardised using training statistics and evaluated on the original scale.

3.2 Methods

We construct three relational graphs representing geographic adjacency, road connectivity, and mobility interaction (Table 1), all defined on the same node set.

Geographic adjacency.

$$A_{ij}^{\text{space}} = \begin{cases} 1, & \text{if zones } i \text{ and } j \text{ share a boundary,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Road connectivity.

$$\begin{aligned} D_{ij}^{\text{road}} &= \min\left\{D_{ij}^{\text{road}}, D_{ij}^{\text{bike}}, D_{ij}^{\text{walk}}, D_{ij}^{\text{road-rail}}, D_{ij}^{\text{walk-rail}}\right\}, \\ A_{ij}^{\text{road}} &= \begin{cases} \dfrac{1}{d_{ij}^{\text{road}} + \epsilon} & d_{ij}^{\text{road}} < \infty, \\ 0 & d_{ij}^{\text{road}} = \infty, \end{cases} \qquad i \neq j. \end{aligned} \tag{2}$$

Mobility interaction.

$$A_{ij}^{\text{mob}} = \frac{f_{ij}}{\sum_j f_{ij}}, \tag{3}$$

Graph normalisation and diffusion operators.

$$A^{(k)} \leftarrow \frac{A^{(k)} + (A^{(k)})^\top}{2},$$

$$L_1^{(k)} = (D^{(k)})^{-1} A^{(k)}, \quad L_2^{(k)} = (D^{(k)})^{-1} (A^{(k)})^\top, \quad L^{(k)} = \frac{L_1^{(k)} + L_2^{(k)}}{2}.$$

Adaptive diffusion fusion.

$$\alpha_k = \text{Softmax}\left(\text{logit}^{(\alpha)}\right)_k, \quad \sum_{k=1}^{K} \alpha_k = 1, \quad L^* = \sum_{k=1}^{K} \alpha_k L^{(k)}.$$

where $\text{logit}^{(\alpha)}$ is a learnable graph scoring vector, α_k denotes the weight assigned to graph view k, K is the number of graph views, $L^{(k)}$ is the graph-specific diffusion operator, and L^* is the fused diffusion operator.

Spatiotemporal modelling.

$$\mathbf{g}_\theta * \mathbf{x} = \sum_{k=0}^{K} \theta_k T_k(L^*)\mathbf{x}, \quad \beta_h = \text{Softmax}\left(\text{logit}^{(\beta)}\right)_h, \quad \sum_{h=1}^{H} \beta_h = 1,$$

$$\mathcal{L} = \sum_{h=1}^{H} \beta_h \frac{1}{N} \sum_{i=1}^{N} \left| y_i^{(h)} - \hat{y}_i^{(h)} \right|.$$

where $T_k(\cdot)$ denotes the k-th Chebyshev polynomial, θ_k is a learnable filter coefficient, β_h is the learned weight assigned to forecast horizon h, and $\mathcal{L}$ is the horizon-weighted MAE loss.

Training and evaluation. Data are split chronologically (70/15/15). Models are trained with Adam [7], early stopping, and gradient clipping, and repeated five times. Performance is evaluated using

$$\text{MAE} = \frac{1}{N}\sum |y - \hat{y}|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum (y - \hat{y})^2}, \quad \text{MAPE} = \frac{100}{N}\sum \left|\frac{y - \hat{y}}{y}\right|.$$

4 Results and Analysis

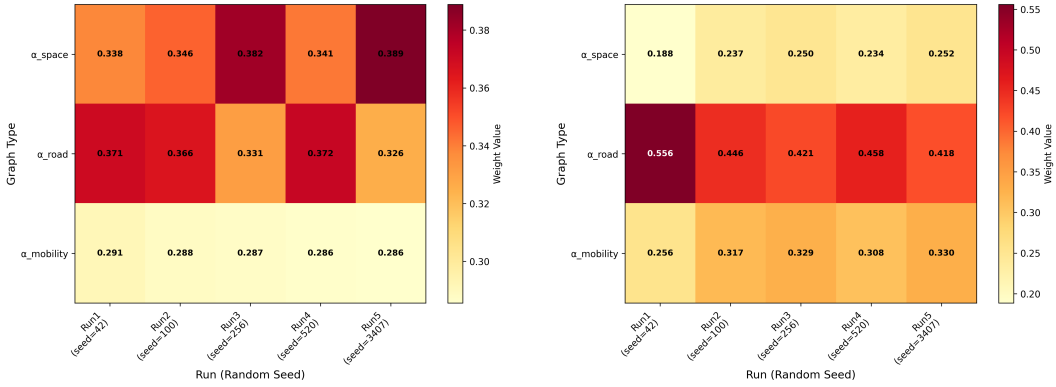
Table 2 summarises forecasting performance in Glasgow and Edinburgh. Across both cities, the road graph provides the strongest single-graph baseline, achieving lower errors than either geographic adjacency or mobility interaction. However, the adaptive fusion model further improves predictive performance, achieving the lowest RMSE and MAPE while maintaining comparable MAE. This indicates that combining multiple spatial perspectives can provide additional predictive benefit when the relative contribution of each relational graph is learned adaptively rather than fixed a priori. In contrast, equal-weight fusion produces substantially higher errors, particularly in Glasgow, suggesting that improvement comes not from graph combination alone, but from how the relational graphs are comparatively weighted.

Table 2 Prediction performance in Glasgow (GLA) and Edinburgh (EDI) (mean ± std over five runs). Lower values indicate better performance.

City	Graph	MAE	RMSE	MAPE (%)
GLA	Road	21.8 ± 0.9	31.5 ± 0.2	68.7 ± 6.2
	Space	23.2 ± 2.0	32.3 ± 1.1	78.4 ± 12.1
	Mobility	22.4 ± 1.5	32.0 ± 0.6	71.6 ± 9.8
	Fusion (Equal weight)	27.4 ± 9.6	36.9 ± 8.9	92.1 ± 49.8
	Fusion Adaptive	**21.8 ± 1.2**	**31.3 ± 0.4↓**	**67.8 ± 8.3↓**
EDI	Road	21.9 ± 1.7	31.8 ± 0.7	73.8 ± 12.7
	Space	23.6 ± 1.9	32.9 ± 0.8	81.2 ± 13.2
	Mobility	22.5 ± 2.6	32.1 ± 1.7	78.4 ± 15.1
	Fusion (Equal weight)	21.3 ± 0.8	33.2 ± 1.7	69.0 ± 4.2
	Fusion Adaptive	**20.9 ± 1.5↓**	**31.6 ± 0.8↓**	**66.0 ± 8.4↓**

Table 3 Learned adaptive spatial weights (α values) in Glasgow (GLA) and Edinburgh (EDI) (mean ± standard deviation over five runs).

Component	GLA	EDI
Road	0.35 ± 0.02	0.46 ± 0.05
Space	0.36 ± 0.02	0.23 ± 0.02
Mobility	0.29 ± 0.00	0.31 ± 0.03

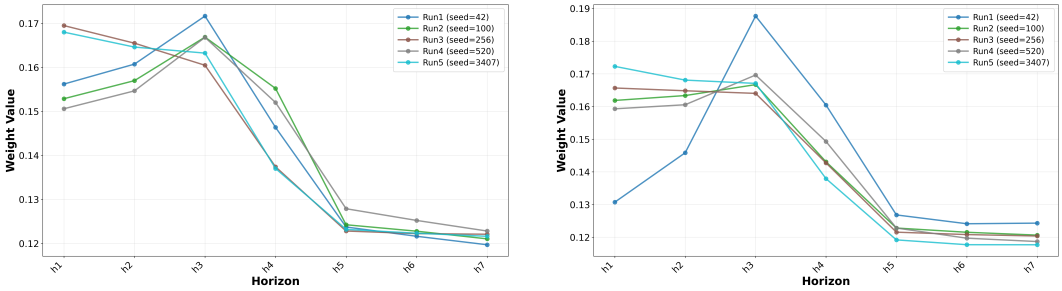


(a) GLA. **(b)** EDI.

Figure 2 Learned spatial weights (α) under the adaptive graph construction in Glasgow and Edinburgh.

The learned spatial weights reveal clear differences between the two cities (Table 3 and Figure 2). In Glasgow, the weighting pattern is relatively balanced: road connectivity (0.35) and geographic adjacency (0.36) contribute at similar levels, while mobility interaction (0.29) plays a smaller role. In Edinburgh, by contrast, road connectivity is more strongly prioritised (0.46), followed by mobility interaction (0.31) and geographic adjacency (0.23). This indicates greater reliance on infrastructural connectivity in Edinburgh, compared with the more balanced relational structure observed in Glasgow.

Although road connectivity provides the strongest individual graph in both cities, the adaptive model assigns non-negligible weights to all three relational graphs. This suggests that no single relational representation is sufficient on its own, and that multiple spatial perspectives contribute to forecasting performance in different ways.



(a) GLA. **(b)** EDI.

Figure 3 Learned temporal weights (β) across prediction horizons for (a) Glasgow and (b) Edinburgh.

Temporal weights (Figure 3) follow a similar pattern in both cities, placing greater emphasis on short- and mid-term horizons and less emphasis on longer horizons. Compared with the spatial weights, which vary across cities, the temporal weighting patterns are more consistent. Taken together, these results suggest that adaptive weighting captures context-dependent spatial structure, while temporal weighting remains comparatively stable across the two urban settings.

5 Discussion and Conclusion

This study suggests that multi-graph spatiotemporal modelling can be understood not only as a technical problem of graph fusion, but also as a problem of spatial perspective selection. By representing multiple relational encodings within a unified framework and learning their relative contributions through adaptive spatial weighting, the proposed approach makes explicit how different relational representations are prioritised within the forecasting process.

The empirical results indicate context-dependent weighting patterns across the two cities. More broadly, the findings suggest that adaptive spatial weighting may provide a useful mechanism for exploring how alternative spatial perspectives contribute to spatiotemporal modelling in different contexts. Rather than identifying a universally optimal spatial representation, the results suggest that the relevance of different relational structures may vary across urban settings.

Understanding the relative importance of spatial perspectives is valuable because different relational graphs imply different explanations of how spatial processes operate. By making these differences explicit, the proposed framework provides interpretative value beyond predictive accuracy and links empirical modelling with broader GIScience questions concerning spatial representation and relational thinking. Its primary contribution therefore lies not in large predictive gains, but in making representational prioritisation observable within a spatiotemporal learning framework and reframing multi-graph learning as a process of selecting among alternative spatial perspectives rather than simply combining multiple graphs.

Although demonstrated using COVID-19 forecasting, the framework may be applicable to a broader range of spatiotemporal phenomena in which multiple relational structures coexist, including traffic forecasting, urban mobility analysis, and environmental monitoring. Nevertheless, the current study is limited to two urban contexts and does not include a full ablation analysis or comparisons with stronger multi-graph baselines. Future work will extend the framework to additional cities, relational layers, and application domains, while further

examining how learned weighting patterns relate to different spatial processes. Overall, this work highlights that spatial representation is not fixed, but can be learned and adapted in a context-sensitive manner within spatiotemporal models.

References

1 Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Chris R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, Matteo Tomasini, and Alessandro Vespignani. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2021. doi:10.1016/j.physrep.2018.01.001.

2 Michael Batty. *The New Science of Cities*. MIT Press, Cambridge, MA, 2013. doi:10.7551/mitpress/9399.001.0001.

3 Stefano Boccaletti, Ginestra Bianconi, Regino Criado, et al. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014. doi:10.1016/j.physrep.2014.07.001.

4 Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022, 2013. doi:10.1103/PhysRevX.3.041022.

5 Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiaowen He. Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1972–1982, 2019.

6 Shengnan Guo, Youfang Lin, Nan Feng, Chuan Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929, 2019. doi:10.1609/aaai.v33i01.3301922.

7 Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. doi:10.48550/arXiv.1412.6980.

8 Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of the Royal Society Interface*, 11(98):20140343, 2014. doi:10.1093/comnet/cnu016.

9 Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2018. doi:10.48550/arXiv.1707.01926.

10 Daniel R. Montello. Scale and multiple psychologies of space. In *Spatial Information Theory: A Theoretical Basis for GIS*, volume 716 of *Lecture Notes in Computer Science*, pages 312–321. Springer, 1993. doi:10.1007/3-540-57207-4_21.

11 OpenStreetMap contributors. Geofabrik download server – scotland extract. https://download.geofabrik.de/europe/united-kingdom/scotland.html, 2023. Accessed 1 March 2023.

12 Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015. doi:10.1103/RevModPhys.87.925.

13 Public Health Scotland. Covid-19 in scotland dataset. https://www.opendata.nhs.scot/dataset/covid-19-in-scotland, 2023. Accessed 1 March 2023.

14 Urban Big Data Centre. Ubdc mobility dataset. https://data.ubdc.ac.uk/datasets/6d87635c-96bb-4b7a-9a2e-676bb250032d, 2023. Accessed 1 March 2023.

15 Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2018. doi:10.48550/arXiv.1710.10903.

16 Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021. doi:10.1109/TNNLS.2020.2978386.

17 Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 753–763, 2020. doi:10.1145/3394486.3403118.

18 Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3634–3640, 2018. doi:10.24963/ijcai.2018/505.

The Same Street, Different Safety Judgments: Evaluative Framing Effects in Daytime and Nighttime Street View Imagery

Tong Qin
Department of Geography, Ghent University, Belgium

Nico Van de Weghe
Department of Geography, Ghent University, Belgium

Wim Fias
Department of Experimental Psychology, Ghent University, Belgium

Haosheng Huang
Department of Geography, Ghent University, Belgium

Abstract

Street View Imagery (SVI) is widely used to study perceived street safety, yet most studies treat safety as a general impression of urban scenes and rely mainly on daytime imagery. It remains unclear whether safety judgments differ across evaluative framings, defined here as the safety question asked of participants, and whether this difference varies between daytime and nighttime scenes. We analyzed 7,320 safety ratings from 61 participants for 120 SVIs, including 60 daytime and 60 nighttime images. Participants judged scenes under two evaluative framings: crime-related safety and traffic-related safety. Using cumulative link mixed models with random intercepts for participants and stimuli, we found that the stimulus set received higher safety ratings in the crime-related framing than in the traffic-related framing. This evaluative-framing effect was stronger for daytime than for nighttime scenes. These findings indicate that perceived street safety in SVI is not a fixed property of image content, but also depends on the safety question asked during judgment.

2012 ACM Subject Classification Information systems → Geographic information systems; Human-centered computing → Empirical studies in HCI; Human-centered computing → User studies

Keywords and phrases street view imagery, spatial cognition, safety perception, evaluative framing, geographic information science

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.32

Category Short Paper

Acknowledgements We want to thank the study participants for their time.

1 Introduction

Urban safety perception shapes how people experience and evaluate cities, including where they move, how they use public space, and how they assess urban quality [5, 6, 16, 9, 15]. From the perspective of spatial cognition, such judgments can be understood as interpretations of spatial information: people not only perceive physical environments of the street, but also assign meaning to them in relation to possible actions, risks, and movement decisions. In recent years, SVI and data-driven methods have made it possible to study perceived safety across large urban areas [13, 2, 3, 4, 7]. This work has shown that visible street features are related to safety judgments and has supported new ways to map, compare, and predict perceived safety across cities and neighborhoods. However, most SVI-based studies treat safety as a general evaluative impression of a street scene, often using daytime imagery as the main visual source [11]. This approach is useful for building comparable datasets and

© Tong Qin, Nico Van de Weghe, Wim Fias, and Haosheng Huang;
licensed under Creative Commons License CC-BY 4.0

models, but it may hide important differences between specific safety concerns. A single general safety score cannot clearly show whether a street feature is related to fear of crime, traffic danger, or a mixture of both. This limits our ability to interpret the relationship between urban environments and perception, and also makes it harder to translate model results into targeted urban design strategies.

Urban safety perception is better understood as a multidimensional judgment rather than a single undifferentiated response [11]. In this paper, evaluative framing refers to the safety question asked of participants, namely crime-related safety or traffic-related safety. We do not assume that crime-related and traffic-related safety are the same type of judgment or rely on the same underlying process; rather, we use them as two distinct safety framings applied to a shared set of street-view images. The same street environment may be evaluated differently when people are asked to consider crime-related safety or traffic-related safety. These evaluative framings may change which visual cues people attend to and how they interpret a scene, and how the scene may support or discourage movement through urban space. This issue may be especially important when comparing daytime and nighttime scenes, because lighting conditions can change the meaning of specific environmental features. For example, greenery is often associated with more positive safety judgments in daytime settings [1], but after dark its effect may weaken or even reverse when vegetation is combined with poor lighting, reduced visibility, or a stronger sense of concealment [12]. It therefore remains unclear whether street-scene safety judgments differ under different evaluative framings, and whether this difference varies between daytime and nighttime scenes. Against this background, we ask two related questions:

- Does evaluative framing change perceived safety across a shared stimulus set?
- Does this evaluative-framing effect differ between daytime and nighttime scenes?

2 Methods

Overview

To examine whether perceived street safety changes with evaluative framing, and whether this effect differs between daytime and nighttime scenes, we conducted a controlled laboratory experiment using SVIs and analyzed the resulting ratings with cumulative link mixed models. The experiment was designed to estimate evaluative-framing effects while holding the stimulus set constant: a shared stimulus set of street scenes was evaluated under two evaluative framings, defined by crime-related and traffic-related safety questions, and both daytime and nighttime scenes were included (Figure 1). Because the outcome variable was an ordered 1-7 safety rating, and because observations were crossed by both participants and stimuli, cumulative link mixed modeling provided an appropriate analytic framework.

Experiment

Sixty-one participants took part in the experiment. The stimuli consisted of 120 SVIs, including 60 daytime scenes and 60 nighttime scenes. Participants were instructed to adopt a pedestrian perspective and rate how safe they would feel in the depicted environment. Safety was evaluated under two evaluative framings: a crime-related safety question and a traffic-related safety question. In the crime-related safety framing, participants rated how safe they would feel with respect to possible criminal events, such as theft, assault, vandalism, robbery, and burglary, while walking as pedestrians. In the traffic-related safety framing, participants rated how safe they would feel with respect to possible traffic accidents while

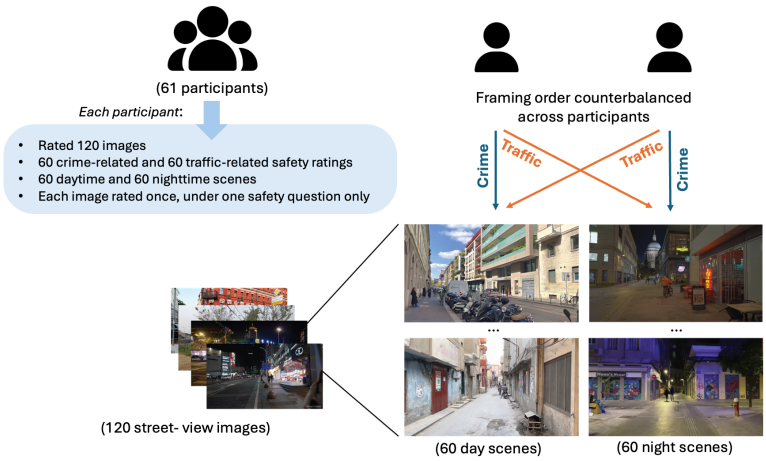


Figure 1 Overview of the experimental design and stimulus assignment. Participants completed both crime-related and traffic-related safety framings, while each image was shown only once to a given participant and under one framing only.

walking as pedestrians. For each trial, participants viewed one street scene and provided a safety judgment on an ordered 1-7 scale, where higher values indicated higher perceived safety. The final dataset contained 7,320 rating observations. Each participant completed both evaluative framings, but each image was shown only once to a given participant and under only one framing. Framing order and image presentation order were counterbalanced across four participant subgroups, reducing repetition and order effects while preserving comparison across the shared stimulus set.

Statistical analysis

The main analyses used cumulative link mixed models (CLMMs), reflecting both the ordinal outcome and the crossed data structure. The dependent variable was an ordered 1-7 safety rating, which has a clear rank order but cannot be assumed to have equal intervals between categories. Simple t-tests or ANOVA would require aggregation or would not adequately account for repeated ratings across both participants and stimuli. Linear mixed models can include random effects, but they treat the rating scale as continuous. CLMMs were therefore more appropriate because they model ordered responses directly while accounting for participant- and stimulus-level variability. We estimated two CLMMs with a logit link and random intercepts for participants and stimuli. Model 1 tested the main effects of evaluative framing and day-night condition while controlling for presentation order. Evaluative framing was coded with crime-related safety as the reference level, and daytime scenes served as the reference level for the day-night factor. Presentation order was mean-centered before analysis.

$$\text{Model 1: } \texttt{rating} \sim \texttt{framing} + \texttt{day_night} + \texttt{order} + (1 \mid \texttt{participant}) + (1 \mid \texttt{stimulus}). \quad (1)$$

Model 2 extended Model 1 by adding the interaction between evaluative framing and day-night condition. This model tested whether the evaluative-framing effect differed between daytime and nighttime scenes.

$$\text{Model 2: } \texttt{rating} \sim \texttt{framing} + \texttt{day_night} + \texttt{framing} : \texttt{day_night} + \texttt{order} + (1 \mid \texttt{participant}) + (1 \mid \texttt{stimulus}). \quad (2)$$

Model interpretation focused on fixed-effect estimates and likelihood-ratio comparisons between nested models. Model 1 assessed whether safety ratings differed by evaluative framing and day-night condition. Model 2 assessed whether the effect of evaluative framing changed between daytime and nighttime scenes. Model fit was evaluated through likelihood-ratio comparisons between nested models. We first compared Model 1 with a reduced model excluding evaluative framing to test the overall contribution of evaluative framing, and then compared Model 2 with Model 1 to test whether adding the evaluative framing × day-night interaction improved model fit. All models were fitted in R using ordinal::clmm with Laplace approximation.

3 Results

Descriptive patterns in observed safety ratings

Figure 2 shows the observed mean safety ratings across evaluative framings and day-night conditions. For daytime scenes, the mean rating was lower in the traffic-related safety framing than in the crime-related safety framing (4.51 vs. 4.79). For nighttime scenes, the same direction was observed, but the difference was smaller (4.46 vs. 4.53). This descriptive pattern suggests that evaluative framing changed perceived safety, with a larger framing-related difference for daytime scenes than for nighttime scenes.

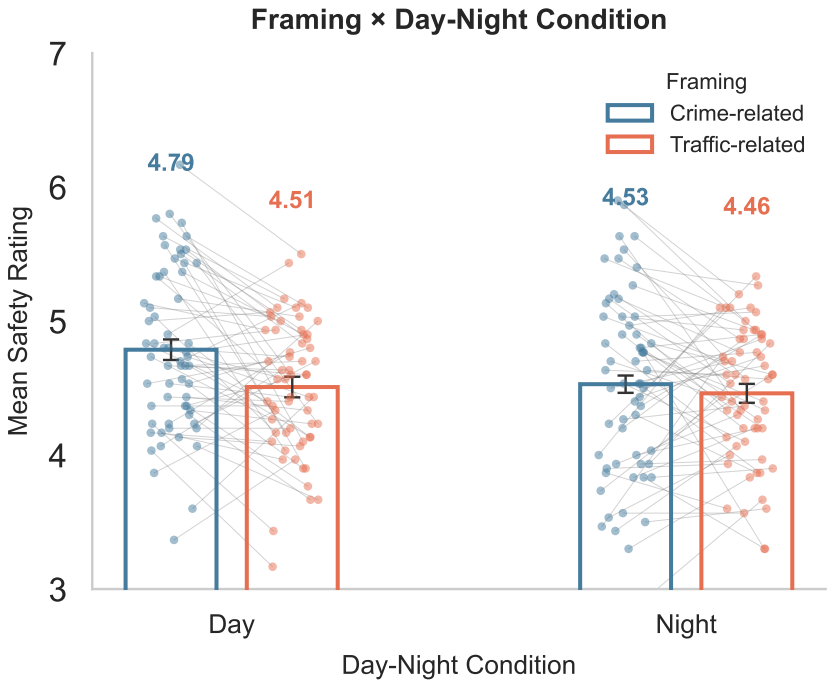


Figure 2 Mean safety ratings by evaluative framing (Crime vs. Traffic) and day-night condition (Day vs. Night).

Evaluative framing changes perceived safety

The CLMM results supported an effect of evaluative framing. In Model 1, ratings were significantly lower in the traffic-related safety framing than in the crime-related safety framing ($b = -0.324$, $SE = 0.043$, $z = -7.576$, $p < .001$). Adding evaluative framing significantly improved model fit compared with a reduced model without evaluative framing, $\chi^2(1) = 57.57$, $p < .001$. In contrast, neither day-night condition nor presentation order showed a significant main effect in this model (night: $b = -0.310$, $p = .344$; presentation order: $b = 0.001$, $p = .750$). These results indicate that, across the shared stimulus set, ratings were lower under traffic-related framing than under crime-related framing.

The evaluative-framing effect differs between daytime and nighttime scenes

Model 2 tested whether the effect of evaluative framing differed between daytime and nighttime scenes. The interaction between evaluative framing and day-night condition was significant (traffic framing × night: $b = 0.439$, $SE = 0.085$, $z = 5.136$, $p < .001$), and the interaction model fit the data better than the main-effects model, $\chi^2(1) = 26.42$, $p < .001$. The main effect of traffic-related framing remained negative in this model ($b = -0.546$, $SE = 0.061$, $z = -8.973$, $p < .001$), while the positive interaction term showed that this negative effect was weaker at night. Neither the main effect of day–night condition nor presentation order was significant (night: $b = -0.529$, $p = .110$; presentation order: $b = 0.001$, $p = .770$). This pattern supports the descriptive result shown in Figure 2: evaluative framing had a stronger effect on daytime scenes than on nighttime scenes.

4 Discussion

This study adds an evaluative-framing perspective to SVI-based research on perceived street safety. Our findings do not contradict content-based approaches, which have shown that visible environmental features are meaningfully related to safety judgments. Rather, they suggest that such approaches remain incomplete if perceived safety is treated as a fixed property of image content alone. The shared set of street scenes received different safety ratings depending on the evaluative framing, that is, the safety question participants were asked to answer. In this sense, perceived safety in SVI is not only content-sensitive but also framing-dependent. The main contribution of this study is to show that the difference between crime-related and traffic-related safety judgments depends on whether street-view images are shown in daytime or nighttime conditions.

A notable finding is that, for the same street scenes, ratings were generally lower under traffic-related framing than under crime-related framing. One possible explanation is that traffic safety judgments rely on more immediate and concrete visual criteria, such as the presence of sidewalks, the clarity of road markings or traffic signs, potential conflicts between vehicles and pedestrians, or the perceived danger of crossings and intersections [8]. Crime-related safety, by contrast, may be evaluated in a more general and diffuse way, because cues related to crime are often less direct and may depend more on broader environmental meaning than on clearly identifiable single-image features [14]. As a result, participants may have applied a stricter evaluative standard under traffic-related framing, leading to systematically lower ratings. However, crime-related and traffic-related safety may involve partly different perceptual and cognitive mechanisms, so the difference between them should not be interpreted as a simple shift along one general safety scale.

At the same time, the evaluative-framing effect was stronger in daytime scenes than in nighttime scenes. The model does not show a uniform day-night shift in safety ratings. Instead, it shows that day-night condition moderated how evaluative framing influenced safety judgments. As shown in Figure 2, ratings were lower in the traffic-related framing than in the crime-related framing for both daytime and nighttime scenes, but the gap between the two framings was larger for daytime scenes. This result suggests that the effect of evaluative framing was not the same across visual conditions.

One possible interpretation is that nighttime scenes provide less clear spatial information for differentiating between types of safety concerns. Reduced visibility, weaker prospect, lower social presence, and a greater possibility of concealment may make crime-related safety judgments more uncertain or more negative [10, 12]. Under such conditions, crime-related safety may move closer to traffic-related safety, reducing the contrast between the two evaluative framings. However, since perceived visibility, uncertainty, and concealment were not directly measured, further work is needed to test this mechanism more directly. Nevertheless, the finding is important for SVI-based research because it shows that day-night condition can change how evaluative framing shapes the interpretation of street-scene information. This matters especially for daytime-dominant SVI research, because the evaluative-framing effect appears to be strongest in the type of imagery on which the field has relied most heavily [11]. The result supports the view that SVI-based safety ratings are not direct readings of image content, but framing-dependent interpretations of spatial information. The meaning of a street feature may depend jointly on the evaluative question being asked and the visual condition under which the scene is represented.

These findings have broader implications for how results from SVI-based safety studies are interpreted and compared. If perceived safety depends partly on evaluative framing, then studies that use different safety questions, or leave the meaning of "safety" underspecified, may not be directly comparable even when they use similar images or modeling pipelines. The results also suggest that future datasets and empirical studies should make the evaluative meaning of safety more explicit. Taken together, these findings challenge the common assumption that perceived safety can be treated as a stable property of street imagery and instead suggest that it emerges from the interaction between visual content and evaluative framing.

5 Conclusion

This study investigated whether perceived street safety varies with evaluative framing and across day-night conditions. The results show that street-scene safety judgments can differ under crime-related and traffic-related safety framings, and that this difference is larger for daytime than for nighttime scenes. These findings suggest that perceived safety in street imagery is not determined by visual content alone but is partly shaped by the safety question asked about the scene. Recognizing this framing dependence is important for interpreting, comparing, and designing SVI-based studies of urban safety perception, especially when safety is otherwise treated as a single general rating.

References

1 Nandita Basu, Oscar Oviedo-Trespalacios, Mark King, Md. Kamruzzaman, and Md. Mazharul Haque. The influence of the built environment on pedestrians' perceptions of attractiveness, safety and security. *Transportation Research Part F: Traffic Psychology and Behaviour*, 87:203–218, 2022. doi:10.1016/j.trf.2022.03.006.

2 Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, 215:104217, 2021. doi:10.1016/j.landurbplan.2021.104217.

3 Zhuangyuan Fan, Fan Zhang, Becky P. Y. Loo, and Carlo Ratti. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences of the United States of America*, 120(27):e2220417120, 2023. doi:10.1073/pnas.2220417120.

4 Yuhao Kang, Jonatan Abraham, Vania Ceccato, Fábio Duarte, Song Gao, Lukas Ljungqvist, Fan Zhang, Per Näsman, and Carlo Ratti. Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in stockholm, sweden. *Landscape and Urban Planning*, 236:104768, 2023. doi:10.1016/j.landurbplan.2023.104768.

5 Eduardo E. Lozano. Visual needs in the urban environment. *Town Planning Review*, 45(4):351–374, 1974. doi:10.3828/tpr.45.4.h43m7270u0m3x968.

6 Jack L. Nasar. Perception, cognition, and evaluation of urban places. In Irwin Altman and Ervin H. Zube, editors, *Public Places and Spaces*, pages 31–56. Springer, Boston, MA, 1989. doi:10.1007/978-1-4684-5601-1_3.

7 Yoshiki Ogawa, Takuya Oki, Chenbo Zhao, Yoshihide Sekimoto, and Chihiro Shimizu. Evaluating the subjective perceptions of streetscapes using street-view images. *Landscape and Urban Planning*, 247:105073, 2024. doi:10.1016/j.landurbplan.2024.105073.

8 Callista A. Ottoni, Joanie Sims-Gould, and Meghan Winters. Safety perceptions of older adults on an urban greenway: Interplay of the social and built environment. *Health & Place*, 70:102605, 2021. doi:10.1016/j.healthplace.2021.102605.

9 Luca Pappalardo, Ed Manley, Vedran Sekara, and Laura Alessandretti. Future directions in human mobility science. *Nature Computational Science*, 3(7):588–600, 2023. doi:10.1038/s43588-023-00469-4.

10 Yunmi Park and Max Garcia. Pedestrian safety perception and urban street settings. *International Journal of Sustainable Transportation*, 14(11):860–871, 2020. doi:10.1080/15568318.2019.1641577.

11 Tong Qin, Nico Van de Weghe, and Haosheng Huang. The role of urban built environments in the perception of safety: A systematic review. *Environmental Impact Assessment Review*, 114:107915, 2025. doi:10.1016/j.eiar.2025.107915.

12 Johan Rahm, Catharina Sternudd, and Maria Johansson. "in the evening, i don't walk in the park": The interplay between street lighting and greenery in perceived safety. *Urban Design International*, 26(1):42–52, 2021. doi:10.1057/s41289-020-00134-6.

13 Philip Salesses, Katja Schechtner, and César A. Hidalgo. The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*, 8(7):e68400, 2013. doi:10.1371/journal.pone.0068400.

14 I. Senna, F. Iglesias, and L. H. Matsunaga. Measuring the effects of Crime Prevention Through Environmental Design (cpted) on fear of crime in public spaces. *Crime Prevention and Community Safety*, 27(1):1–17, 2025. doi:10.1057/s41300-025-00223-0.

15 Petrus te Braak and Theun Pieter van Tienoven. Navigating the city: A systematic literature review on women's perceived safety in urban public space. *Cities*, 162:105907, 2025. doi:10.1016/j.cities.2025.105907.

16 Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H. Fung, Hui Lin, and Carlo Ratti. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160, 2018. doi:10.1016/j.landurbplan.2018.08.020.

Decision-Making Under Uncertainty in Tsunami Evacuation: Individual Differences in Information Consultation and Risk Perception

Shota Iwama
Tokyo University of Science, Japan

Sihwan Lee
Tokyo University of Science, Japan

Hong Hu
Tokyo University of Science, Japan

Takanori Nanahara
Nagoya University, Japan

Abstract

In tsunami evacuation, the two fundamental strategies are horizontal evacuation (moving away from the coastline toward a safer inland location) and vertical evacuation (moving to the upper floors of a building). However, few studies have investigated how individuals make decisions when both strategies are simultaneously available. This study examines how individual differences in lead-time information consultation and in degree of risk aversion affect strategy selection when evacuation begins from inside a high-rise building. A virtual reality (VR) experiment was conducted in an urban environment with 88 participants. Participants evacuated from a seven-story building under three location conditions: Near the Sea, Intermediate, and Near the Mountain. The results revealed that in the absence of lead-time information consultation, participants predominantly opted for vertical evacuation, relying primarily on the perceived safety afforded by the building's elevation. In contrast, when participants consulted the lead-time information, they tended to select horizontal evacuation. In addition, participants with a higher degree of risk aversion were more likely to select vertical evacuation, but only in the Intermediate condition, where the environmental cues were most ambiguous. These results suggest that individual differences (i.e., information consultation and risk perception) play an important role in evacuees' decision-making. Overall, the findings indicate that providing personalized information suited to evacuees' situations can reduce the influence of individual differences and support safer evacuation decisions.

2012 ACM Subject Classification Information systems → Spatial-temporal systems

Keywords and phrases Tsunami, Evacuation Strategy, Information Consultation, Risk Perception, Decision-Making, Virtual Reality

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.33

Category Short Paper

Supplementary Material *Dataset*: `https://osf.io/3uqv8/overview`

Funding This study was supported by the Obayashi Foundation (Grant No. 2025 Research-36-149).

Acknowledgements This study was approved by the Ethics Committee of Tokyo University of Science (No. 25027). The authors thank the Department of Architecture, Faculty of Engineering, Tokyo University of Science, for their technical support and for providing the research facilities.

© Shota Iwama, Sihwan Lee, Hong Hu, and Takanori Nanahara;
licensed under Creative Commons License CC-BY 4.0
17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 33; pp. 33:1–33:8

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

1.1 Uncertainty in Tsunami Evacuation Strategy Selection and Individual Differences

In tsunami evacuation, selecting a safer destination is crucial to survival. Horizontal evacuation and vertical evacuation are the two fundamental strategies. Horizontal evacuation involves moving away from the coastline to a safer inland location, while vertical evacuation involves moving to the upper floors of a building. While staying in a high-rise building avoids the immediate risk of tsunami submersion, it may entail long-term risks such as structural collapse or isolation (e.g., shortage of relief supplies). Therefore, if time permits, moving to a safe inland area through horizontal evacuation may be considered the safest option.

From a motivational perspective, people who select horizontal evacuation tend to question the structural safety of tsunami vertical evacuation buildings and believe that inland areas can be reached before the tsunami arrives. On the other hand, those who opt for vertical evacuation tend to estimate that reaching inland areas before the tsunami arrival is difficult [4]. Generally, horizontal evacuation involves extended travel times but provides higher destination safety. Conversely, vertical evacuation requires minimal travel time but presents greater ambiguity regarding long-term safety.

Evacuees' geographic location (e.g., coast vs. inland) may significantly affect their perception of inland reachability. However, when geographical cues are ambiguous – for example, when evacuees are located at an intermediate distance from both the coast and high ground – they must make a decision under two uncertainties: the possibility of reaching the destination before the tsunami arrival ("reachability") and the safety of the destination ("safety"). Under such uncertain circumstances, individual differences may play an important role in their decision-making. Reachability can be evaluated by comparing the time required to reach the evacuation site (lead-time) and the tsunami arrival time [8]. Individual differences in the use of available evacuation information may influence how evacuees cope with this uncertainty. Indeed, reviewing the evacuation plan on the floor map has been shown to influence subsequent route selection during an emergency [7]. In addition, evacuees' degree of risk aversion may affect their decision-making, as the two strategies involve different types of risk [6]. However, the influence of these individual factors on tsunami evacuation strategy selection remains poorly understood.

1.2 Study Objectives

This study aims to address two questions in tsunami evacuation. First, it examines how multiple uncertainties and the trade-offs between the two strategies affect evacuees' strategy selection. Second, it investigates how individual differences in information consultation and risk perception affect strategy selection. To address these questions, an experiment was conducted in a desktop-based virtual reality (VR) environment, in which participants were required to evacuate in response to a tsunami under several scenarios. The experiment assumes evacuation from a high-rise building, where the building's height already provides a certain level of safety. In this context, this study investigated how consulting lead-time information and participants' degree of risk aversion affect the selection between staying in the building (vertical evacuation) and moving to a distant inland area (horizontal evacuation). Furthermore, from the perspective of behavioral geography, this study discussed how participants interpreted the available information and how these interpretations influenced their final strategy selection.

2 Methods

2.1 Participants and Materials

Data were collected from 88 participants aged 19 to 64 years (42 males and 46 females; M_{age} = 38.9 years; SD = 9.95) recruited through an online platform. All participants were native Japanese speakers, and the experiment was conducted in Japanese.

A virtual city was designed as a coastal city between the sea and the mountains (Figure 1). It was modeled using Rhinoceros 8 and imported into Unity, which was used to implement the experimental software. Each participant completed the experiment remotely using their own personal computer, without direct in-person supervision by the researchers. They moved within the environment using the keyboard (WASD keys) and controlled the view using the mouse. Movement speed was limited to walking (2.0 m/s) and running (4.3 m/s). Participants' position and viewing direction were recorded at 10 fps.

2.2 Experimental design

An area of high ground on the mountainside served as the destination for horizontal evacuation. Three high-rise buildings were placed at different locations: Near the Sea, Intermediate, and Near the Mountain (Figure 1). These target buildings were seven stories tall and high enough to remain safe from tsunami flooding. In each trial, the tsunami evacuation began inside one of them, and the rooftop of that building served as the destination for vertical evacuation.

A map of the city was displayed on the wall of the room where the evacuation began, and participants were free to consult it (Figure 2). The map provided the lead-time that served as critical temporal information regarding destination reachability, along with a simplified bird's-eye view of the city, the participant's current location (i.e., the location of the building), a marker for the evacuation site on the high ground, and an arrow indicating the route. While the primary function of a map is to provide spatial information, we designed the virtual environment with a simple grid-like structure and included a practice phase to familiarize participants with the environment beforehand. Consequently, participants were expected to have sufficient spatial knowledge, allowing the map to function primarily as a source of temporal information. This enabled us to analyze whether consulting the lead-time information via the map influenced participants' evacuation strategy selection.

Participants were instructed to complete the evacuation before the tsunami's arrival time, which was displayed on the screen at the beginning of the evacuation. To equalize the temporal difficulty of evacuation across all location conditions (i.e., to avoid a situation where horizontal evacuation would be easier when starting from the Near the Mountain condition), the tsunami arrival time was adjusted for each condition so that the lead-time margin (the tsunami arrival time - the lead-time) was fixed at 2 minutes.

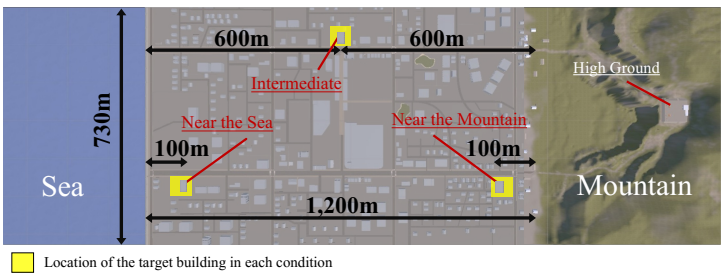


Figure 1 Environmental layout and locations of the evacuation buildings under three conditions.

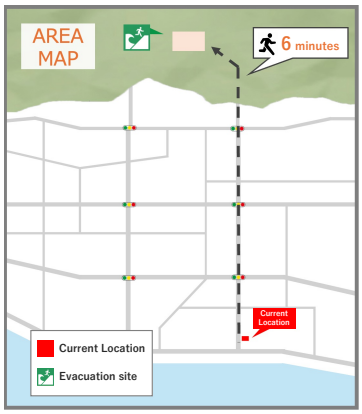


Figure 2 Map of the city (Near the Sea condition). The original text on the map was in Japanese and has been translated into English.

Table 1 Location conditions and evacuation timing parameters.

Location Condition	Tsunami Arrival Time [min]	Lead-Time [min]	Lead-Time Margin [min]
Near the Sea	8	6	2
Intermediate	6	4	2
Near the Mountain	4	2	2

2.3 Procedure

The experiment consisted of three phases: practice, main task, and questionnaire survey.

Practice. First, participants completed an item collection task in the same virtual city used in the main task. During this task, they collected capsule-shaped objects placed throughout the city one by one. Through this task, participants familiarized themselves with the movement speed and the scale of the city. They also gained prior knowledge about the structure of the city, including the two possible evacuation sites, i.e., the high ground and the rooftops of the target buildings, that would appear in the subsequent main task. Additionally, task completion times were recorded.

Main task. The main task was repeated three times, once for each location condition shown in Figure 1. In each trial, participants started from a location near the target building assigned to that condition. They were guided to enter the building and instructed to complete a dummy task requiring them to head to a designated room on the seventh floor and open the window. When they entered the room, a map showing lead-time information was visible on the opposite wall, but they were not explicitly instructed to consult it. Immediately after completing the dummy task, an earthquake occurred, represented by a shaking view on the screen and the sound of the earthquake warning alert. Once the shaking subsided, a major tsunami warning was issued, after which the tsunami arrival time was displayed on the screen, and participants were instructed to evacuate to a safer location. Participants then began moving toward the evacuation destination they considered safest. Each trial terminated when the participant completed the evacuation or when the tsunami arrival time had elapsed. The main task ended after trials in all location conditions were completed. To avoid order and learning effects across location conditions, the order of the trials was randomized for each participant.

Questionnaire survey. After the main task, participants completed a questionnaire. It assessed demographic information (age and gender); experience with and knowledge of tsunamis; usability of the movement controls and perceived realism of the experimental environment (both rated on a five-point Likert scale); and self-reported behavioral aspects during the experiment, such as their reasons for strategy selection and whether they had consulted the map. In addition, participants completed a binary-choice task designed to measure their degree of risk aversion [1], which is based on expected utility theory.

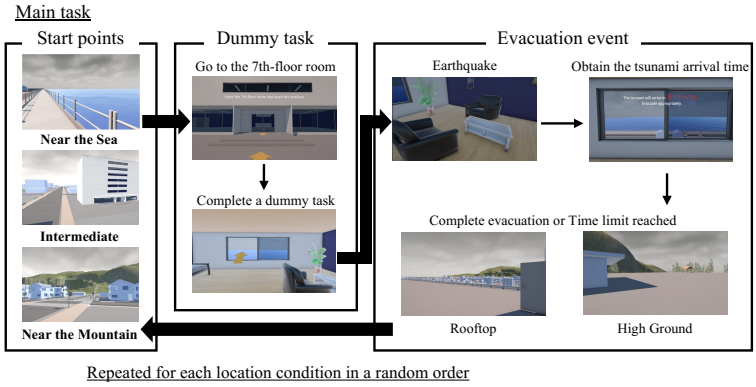


Figure 3 Workflow of the main task.

2.4 Analysis

Evacuation Strategy Selection Rates. For each location condition, the proportions of horizontal and vertical evacuation strategy selections were compared using binomial tests. To evaluate the effect of information consultation, trials were categorized into two groups based on whether participants consulted the map or not, as determined by the duration of the virtual camera orientation toward the map exceeding a predefined threshold and questionnaire responses.

Generalized Linear Models. In addition, generalized linear models (GLMs) for each location condition were constructed to examine the effects of the multiple factors. The response variable was the selected evacuation strategy, with horizontal evacuation coded as 1 and vertical evacuation coded as 0. Explanatory variables included map consultation and risk aversion, the central concerns of the study, along with demographic factors known to be associated with navigation behavior, namely age and gender (male = 1). The completion time of the item collection task was also included as an index of operational proficiency in movement control in the virtual environment, with shorter times indicating higher skill.

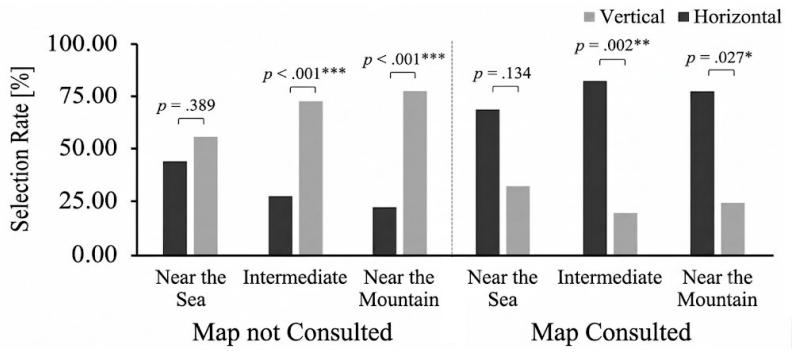
3 Results and Discussion

3.1 Trends in Evacuation Strategy Selection Rates

Figure 4 shows the results of the binomial tests. Participants who did not consult the map selected vertical evacuation significantly more often in the Intermediate and the Near the Mountain conditions (both $p < .001$). Without the lead-time information presented on the map, participants had to estimate the travel time based only on their own sense of distance. They likely relied on the building's height, which was more obvious than the time

needed to evacuate horizontally. Indeed, "building height" was the most frequently reported reason for selecting vertical evacuation in the questionnaire, which is consistent with previous research [3]. Additionally, participants were already at a high elevation when the evacuation began. This may have induced status quo bias [5] and loss aversion [2], which may have deterred participants from moving outside at the cost of their current perceived safety.

In contrast, participants who consulted the map selected horizontal evacuation significantly more often in the Intermediate and the Near the Mountain conditions ($p = .002$ and .027, respectively). By comparing the lead-time information with the tsunami arrival time, they were able to estimate the lead-time margin. This suggests that such objective information effectively reduced the uncertainty regarding reachability and may have replaced the subjective assessment of reachability. In other words, even when the building they are in offers a certain level of safety due to its elevation, people may select a safer evacuation destination if they can obtain clear and reliable information about the time margin.



* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 4 Results of binomial tests for evacuation strategy selection rates.

3.2 Effects of Individual Differences on Strategy Selection

Table 2 summarizes the fitted GLMs. The effect of map consultation was significant in the Intermediate and Near the Mountain conditions (both $p < .001$), which is consistent with the results of the binomial tests. Although the effect was not significant in the Near the Sea condition, a weak tendency was observed ($p = .058$). In this condition, the imminent visual and geographical threat of being very close to the sea may have had a more dominant influence on participants' decision-making than objective numerical information (i.e., lead-time on the map). These results suggest that consulting the lead-time information may promote horizontal evacuation across location conditions. However, this study did not collect qualitative data on participants' reasoning; therefore, we cannot determine why participants chose whether to consult the map. Future studies should investigate the factors underlying map-consultation behavior by collecting open-ended responses about participants' decision-making processes.

In addition, participants with a higher degree of risk aversion were significantly more likely to select vertical evacuation. This effect was observed exclusively in the Intermediate condition ($p = .040$). In the Near the Sea and Near the Mountain conditions, geographic factors, namely the distance to the sea and to the high ground, may have had a stronger influence on decision-making. In contrast, in the Intermediate condition, where such geographic factors were weaker, individual differences may have played a more important role in decision-making.

Furthermore, older participants selected vertical evacuation significantly more often in the Intermediate and the Near the Mountain conditions ($p = .032$ and .044, respectively). This may reflect older participants' greater concerns about the physical demands of horizontal evacuation, including longer travel distance and time, as well as a fear of struggling to navigate the VR environment due to a lack of gaming experience.

Table 2 Summary of the fitted GLMs for evacuation strategy selection.

Location Condition	Variable	β	SE	OR	p
Near the Sea	(Intercept)	2.902	1.504	18.215	0.054
	Map consultation	1.080	0.569	2.946	0.058
	Risk aversion	-0.202	0.120	0.817	0.092
	Age	-0.040	0.024	0.961	0.101
	Gender (male = 1)	0.742	0.488	2.100	0.129
	Operational proficiency	-0.056	0.089	0.946	0.530
Intermediate	(Intercept)	3.528	1.713	34.041	0.040^{*}
	Map consultation	2.702	0.671	14.903	$< 0.001^{***}$
	Risk aversion	-0.297	0.145	0.743	0.040^{*}
	Age	-0.061	0.028	0.941	0.032^{*}
	Gender (male = 1)	0.518	0.558	1.678	0.354
	Operational proficiency	-0.050	0.101	0.951	0.618
Near the Mountain	(Intercept)	1.897	1.701	6.665	0.265
	Map consultation	2.496	0.659	12.129	$< 0.001^{***}$
	Risk aversion	0.038	0.139	1.039	0.784
	Age	-0.061	0.030	0.941	0.044^{*}
	Gender (male = 1)	0.743	0.555	2.102	0.181
	Operational proficiency	-0.130	0.106	0.878	0.219

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4 Conclusion

This study examined the selection of tsunami evacuation strategies when evacuation begins from a high-rise building by conducting a desktop-VR experiment. The results showed that participants who did not consult the map tended to select vertical evacuation by relying on the elevation of the building. In contrast, when they were able to estimate the lead-time margin from the temporal information, horizontal evacuation was selected more often, across all location conditions. In addition, individuals' degree of risk aversion affected strategy selection when the environmental cues were ambiguous (i.e., the Intermediate condition).

These findings indicate that individual differences, including information consultation and risk perception, may play an important role in tsunami evacuation decision-making under uncertainty. This suggests that some individuals may make unsafe decisions. In this context, providing information may effectively mitigate the influence of individual differences and encourage safer evacuation strategies. For example, to promote vertical evacuation, it may be useful to emphasize information about safety, such as building height. To promote horizontal evacuation, it may be useful to emphasize information that helps people to assess reachability, such as tsunami arrival time and lead-time. Since such information depends on each evacuee's location and situation, mobile devices and GPS offer significant potential for delivering personalized information. Particularly in spaces with high uncertainty, as represented by the Intermediate condition in the present study, providing objective spatiotemporal information

plays a crucial role in correcting biases in individuals' subjective risk perception. However, because this study used a controlled VR environment, a more comprehensive analysis is needed before applying these findings to real-world environments.

References

1 Charles A. Holt and Susan K. Laury. Risk Aversion and Incentive Effects. *The American Economic Review*, 92(5):1644–1655, 2002.

2 Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, 1979. doi:10.2307/1914185.

3 Jorge León, Alejandra Gubler, Patricio Catalán, Matías Correa, Javiera Castañeda, Gianni Beninati, and Alonso Ogueda. Assessing potential tsunami vertical-evacuation practices: A study of four cases in Chile using virtual reality and GIS. *International Journal of Disaster Risk Reduction*, 98:104098, November 2023. doi:10.1016/j.ijdrr.2023.104098.

4 Jamie W. McCaughey, Ibnu Mundir, Patrick Daly, Saiful Mahdi, and Anthony Patt. Trust and distrust of tsunami vertical evacuation buildings: Extending protection motivation theory to examine choices under social influence. *International Journal of Disaster Risk Reduction*, 24:462–473, September 2017. doi:10.1016/j.ijdrr.2017.06.016.

5 William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, March 1988. doi:10.1007/BF00055564.

6 Jiguang Shi, Ning Ding, Hao Wang, and Yang Wang. How risk preference affects evacuees' route choice in buildings: An IVR-based experimental study. *Safety Science*, 187:106840, July 2025. doi:10.1016/j.ssci.2025.106840.

7 Dajana Snopková, Pavel Ugwitz, Zdeněk Stachoň, Jiří Hladík, Vojtěch Juřík, Ondřej Kvarda, and Kubíček Petr. Retracing evacuation strategy: A virtual reality game-based investigation into the influence of building's spatial configuration in an emergency. *Spatial Cognition & Computation*, 22(1-2):30–50, April 2022. doi:10.1080/13875868.2021.1913497.

8 Nam Yi Yun and Masanori Hamada. Evacuation Behavior and Fatality Rate during the 2011 Tohoku-Oki Earthquake and Tsunami. *Earthquake Spectra*, 31:pp. 1237–1265, August 2015. doi:10.1193/082013EQS234M.

Simulating Path Integration with Continuous Attractor Neural Networks

Yulin Wu
Shenzhen Graduate School, Peking University, Guangdong, China
University of Leeds, UK

Ed Manley
University of Leeds, UK

Abstract

This paper explores how grid cell-inspired neural dynamics can simulate path integration during human navigation. We combine smartphone GPS traces, trajectory interpolation, and a two-dimensional continuous attractor neural network (2D CANN) to model path integration in York and Leeds, United Kingdom. The model reproduces grid cell-like firing and estimates trajectories that closely match mapped movement. We further introduce a logarithmic psychophysical transformation of velocity to examine how perceived speed may distort cognitive spatial structure, especially under mixed transportation modes. The results suggest that velocity-driven path integration can computationally account for asymmetric distance estimation and cognitive collage-like distortions in urban environments.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning; Computing methodologies → Bio-inspired approaches

Keywords and phrases cognitive map, grid cell, continuous attractor neural network, path integration, spatial cognition, urban mobility

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.34

Category Short Paper

1 Introduction

The metric structure of spatial mental representation, and the Euclidean form built upon it, is central to cognitive geography. The status of Euclidean space in human spatial cognition has long motivated debate [16, 7, 18]. Researchers generally agree that cognitive space differs from physical space and often contains biases or distortions. Geographers and urban analysts have examined associations between geodesic and cognitive distance [15], as well as built-environment features that influence perceived distance or accessibility [12]. Recent findings question the relevance of Euclidean representation, highlighting the role of topological structures [3] and non-integrated cognitive collages [17].

Cognitive neuroscience has provided mechanistic evidence for the formation of Euclidean cognitive space. One of the most influential findings is the discovery of *grid cells* in the medial entorhinal cortex (MEC). Grid cells are activated at locations that tessellate Euclidean space in a hexagonal lattice. This firing pattern represents metric distance, Euclidean spatial relations, and coarse direction [8]. Subsequent work has shown that grid cells encode important properties of Euclidean space because their activity is driven by velocity signals. As an individual moves continuously, grid cells integrate velocity inputs and iteratively update their encoding of current position. This process, analogous to SLAM or inertial navigation, is known as *path integration* [5].

Path integration is highly relevant to cognitive maps in human geography and urban science. Grid cells in the human brain may register travel speed during everyday mobility and, through path integration, partially shape spatial layout understanding. Perceived travel

© Yulin Wu and Ed Manley;
licensed under Creative Commons License CC-BY 4.0

speed is influenced by transportation mode, attention, and individual differences [13, 21]. Estimating cognitive spatial relationships consistent with principles of spatial cognition would be meaningful for spatial theory in GIS and for human-centered geospatial applications. Despite this importance, relatively few studies have explored this topic.

Recent computational neuroscience now provides toolkits for simulating multiple neural cell types, including grid cells. By constructing a recurrent neural network, inputting signals such as velocity, and iteratively simulating the dynamic system until it reaches a stable state, specific firing patterns can emerge. This recurrent dynamic system is an attractor neural network (ANN), and the grid-cell system is commonly modeled as a continuous variant, namely a continuous attractor neural network (CANN). We use these tools to model grid cells and path integration, and to examine how these neural processes may systematically distort cognitive maps.

2 Methods

This initial study uses GPS trajectories recorded by a smartphone during explorations of York and Leeds, two major cities in northern England. The trajectories covered urban and rural areas of both cities. GPS positions were recorded with the smartphone application *FootPrint* and exported as time series of geolocated coordinates. Time was recorded as local timestamps, and the application-reported location error averaged 8.94 m (SD = 7.89 m). York is smaller than Leeds and its attractions are more spatially concentrated. Trajectories combine walking and bus travel.

Mobility in urban environments includes at least two forms: staying at a place for an activity and moving toward the next place [14]. Staying may occupy much of the recording period without movement. We removed stationary periods to reduce data volume and accelerate the neurodynamic simulation. Because users do not generate dense moving points while stationary, we calculated time differences between consecutive GPS records. The 90th percentile was used as a threshold, and longer gaps were compressed to it. After compression, the trajectories approximate continuous movements without stays. The mean time difference between consecutive GPS records was 74.43 seconds (SD = 113.89 s), still too coarse for precise velocity estimation. We therefore used `RatInABox` [6] to interpolate intermediate points between adjacent GPS records at 0.02 s resolution using cubic splines, and then estimated the corresponding velocity time series. This fine temporal resolution is required for numerical stability of the CANN differential-equation integration; the spline interpolation produces smooth velocity signals between GPS fixes while keeping the overall trajectory shape anchored to the actual GPS data. The resulting trajectories have high spatiotemporal resolution and dense sampling, with 90,173 records for York and 133,485 for Leeds.

To model path integration supported by grid cells, we use the computational neuroscience toolbox `CANNs` to simulate a continuous attractor neural network that generates grid-cell-like firing [10]. We briefly introduce the continuous ANN theory of grid cells here. Although researchers generally agree that grid-cell firing is driven by velocity input and neural encoding of current position, the mechanism by which velocity produces a hexagonal firing pattern remains debated. One prominent theory argues that this pattern emerges from an internally interconnected recurrent neural network. Neurons send excitatory or inhibitory signals to one another, and as the system evolves iteratively, cell activation reaches a stable dynamic equilibrium. If the system deviates from this equilibrium, internal interactions drive it toward a stable state, which may or may not match the previous one. Mathematically, the

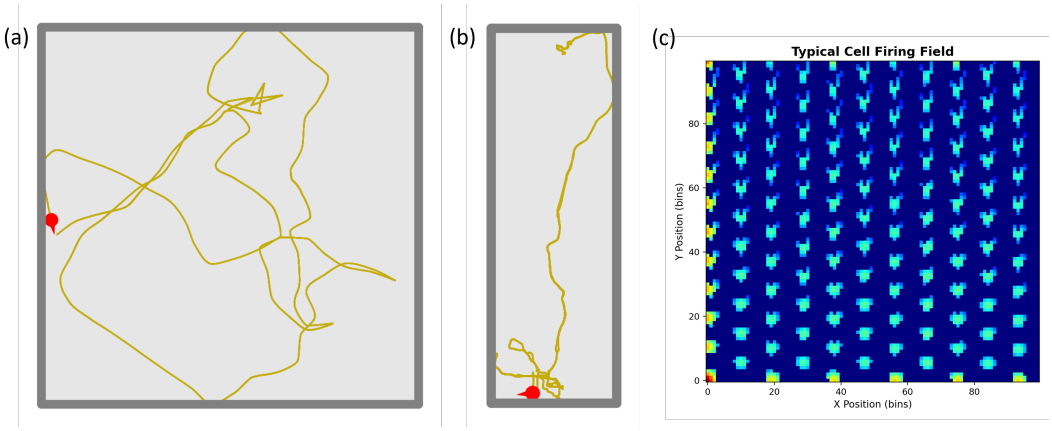


Figure 1 Interpolated trajectories in York (a) and Leeds (b), displayed as if they occurred in an experimental arena. Panel (c) shows the simulated firing pattern of a single grid cell in the CANN system defined in this study, generated in a controlled square arena (side length 2.2 m, movement speed 0.3 m/s). Colour intensity indicates firing rate (brighter means stronger activation); the distinct peaks form the canonical hexagonal lattice.

Table 1 Important parameters used in the CANN grid-cell simulation.

Parameter	Value	Explanation
τ	0.01	Time constant controlling response and decay of neural activity.
α	0.10	Coupling strength between velocity input and activity-bump displacement.
W_ℓ	3.0	Spatial scale of local recurrent connectivity.
λ_{net}	17.0	Internal spatial periodicity controlling the grid-like firing scale.

neural network forms a dynamic system, and stable states are solutions to the corresponding differential equations. The set of all stable solutions is called an attractor. External inputs can break the equilibrium by injecting directed excitatory or inhibitory signals, after which the system forms a new stable state. Computational neuroscientists therefore model grid cells as a neural network with attractor dynamics, that is, as an ANN [11]. In terms of neural connectivity, this ANN forms a torus-like structure. Neurons are arranged as a square sheet in which each cell excites nearby cells and inhibits more distant cells [2], while opposite edges are treated as adjacent. When mammals move through an environment, the excitatory and inhibitory signals shift directionally, changing cell activation. As a result, each neuron in the sheet can generate a hexagonal lattice of spatial activation.

As a representative two-dimensional continuous attractor neural network (2D CANN), the grid-cell system can be modeled with `CANNs`, a library for simulating attractor neural networks and their time-dependent evolution. We define a neural sheet with 40×40 cells. To obtain stable results, we set the main parameters according to official guidance [10], as summarized in Table 1. The 2D CANN tracks the firing state of each cell during movement and decodes an inferred trajectory from neural activity. We treat this trajectory as a computational representation of the cognitive map.

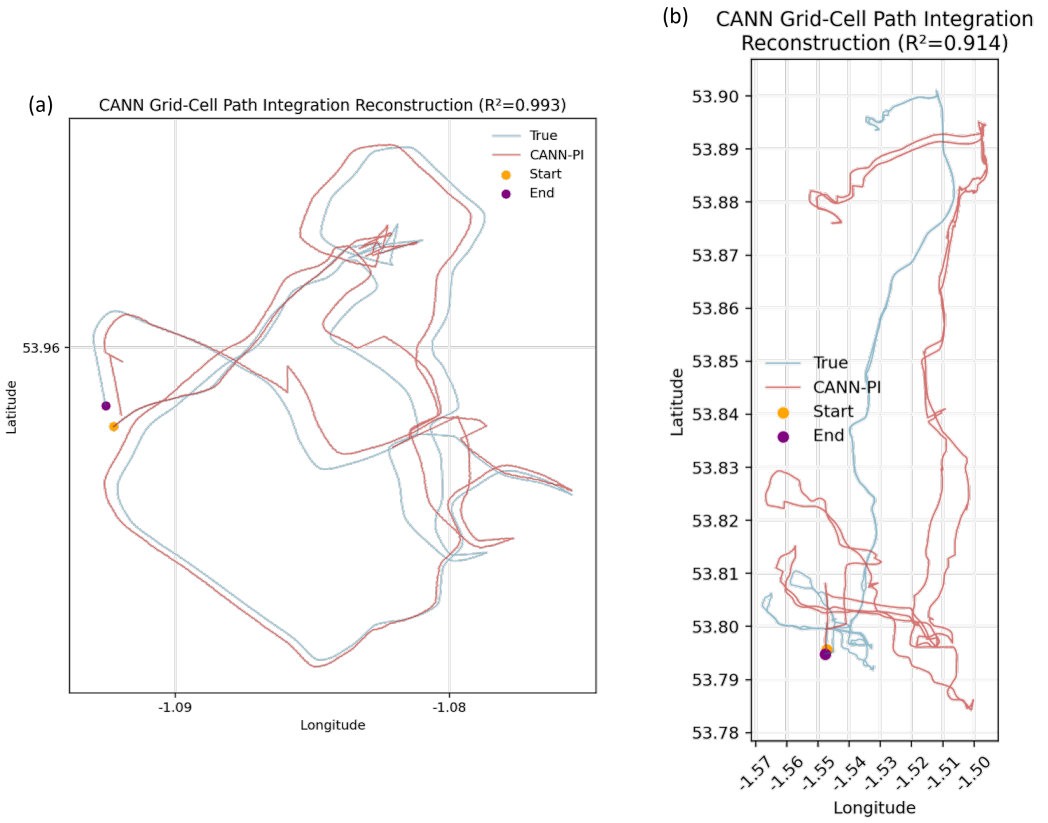


Figure 2 Path integration results of the CANN grid-cell system in York (a) and Leeds (b).

3 Results

The grid-cell network defined as a 2D CANN reproduces the canonical grid-cell firing pattern and supports path integration. The simulated grid cell shows the typical hexagonal lattice firing pattern (Figure 1c). When the inferred trajectory is decoded from the network and compared with mapped mobility, the system behaves like inertial navigation. It delineates the displacement vector from the starting point with high accuracy in York (Figure 2a). To quantify trajectory similarity, we compute cumulative displacement coordinates from the displacement increment series (Δx, Δy) of both the decoded and actual trajectories, then regress the decoded coordinates against the actual coordinates (x and y pooled) using univariate linear regression. This captures shape fidelity while remaining insensitive to overall scale differences, which is appropriate because the grid-cell path-integration system is primarily concerned with relative spatial relationships between locations rather than absolute metric distances. The resulting R^2 scores are 0.994 in York and 0.914 in Leeds. Notably, the estimated Leeds trajectory is distorted: round trips between the city center and Harewood House are slightly shortened. This suggests that the grid-cell CANN system itself can contribute to spatial cognitive distortion.

We further apply this tool to explore potential cognitive-map distortions. A typical question is whether motorized transportation distorts the metric structure of an individual's cognitive space. We hypothesize that such an effect exists and mainly arises from biased speed estimation. Prior research shows that transportation mode and vehicle type can bias

perceived movement speed [13]. Here, we adopt the Weber-Fechner law, a psychophysical rule of stimulus perception, to relate actual and perceived movement speed through a logarithmic function:

$$v_p = k \ln(v_a), \tag{1}$$

where

$$k = \frac{v_b}{\ln(v_b)}. \tag{2}$$

Here, v_p and v_a denote perceived and actual velocity, respectively. The mapping depends on coefficient k, which depends on a user-defined velocity benchmark v_b. The second equation implies that individuals estimate the benchmark velocity relatively accurately. We set $v_b = 5$ m/s, approximately 18 km/h, or cycling speed. This assumption is reasonable for the present exploratory analysis. For slower activities such as walking or running, higher speed usually implies greater physical effort, so people may overestimate movement speed. Conversely, in faster modes such as buses, private cars, or trains, people may not fully perceive speed because of attention, restricted visual exposure, or being inside an enclosed vehicle. They may therefore underestimate travel speed.

We transform all actual instantaneous velocities into perceived velocities with this logarithmic function and input the biased signals into the CANN dynamic system. The result is shown in Figure 3. Because visitors in York only walked through the city and their speed was relatively stable, the path-integration trajectory is slightly larger than the real trajectory because walking speed is overestimated. Nevertheless, its overall shape remains highly consistent with the mapped trajectory, with a squared R score of 0.942 (Figure 3a). The Leeds case differs because mobility included bus travel. Distances between the origins and destinations of bus trips are visibly shrunk (Figure 3b). Movements within Harewood House Trust and downtown Leeds are disproportionately enlarged, resembling "magnifying glass" bias in cognitive map distortion [4]. In this sense, the cognitive map simulated by CANN neural dynamics and a simple psychophysical rule may be closer to subjective spatial cognition than to geomatic survey space. Another important finding is that the real Leeds trajectory starts and ends at the train station, as indicated by the orange and purple points in Figure 3b. In the estimated trajectory, however, the two points are far apart. One reason is that the outbound trip from downtown to the University was on foot, whereas the return trip was by bus. This difference produced asymmetric distance estimation between the two locations, causing the estimated endpoint to deviate from the starting point. This representation of asymmetric distance estimation and positional ambiguity captures an important aspect of the cognitive collage phenomenon [17].

4 Discussion

This study introduces computational neuroscience theories and approaches to model grid cells and their tracking of locomotion. The simulations demonstrate the model's capacity to perform path integration and compute movement-based spatial estimates. A central contribution is that a simple psychological assumption about speed perception can produce a cognitive mapping representation broadly consistent with spatial cognitive intuition. The logarithmic function used here may oversimplify distance perception, and the cycling-speed benchmark is partly arbitrary. Nevertheless, with more systematic quantitative study of human distance perception, this approach may support more plausible computational cognitive maps.

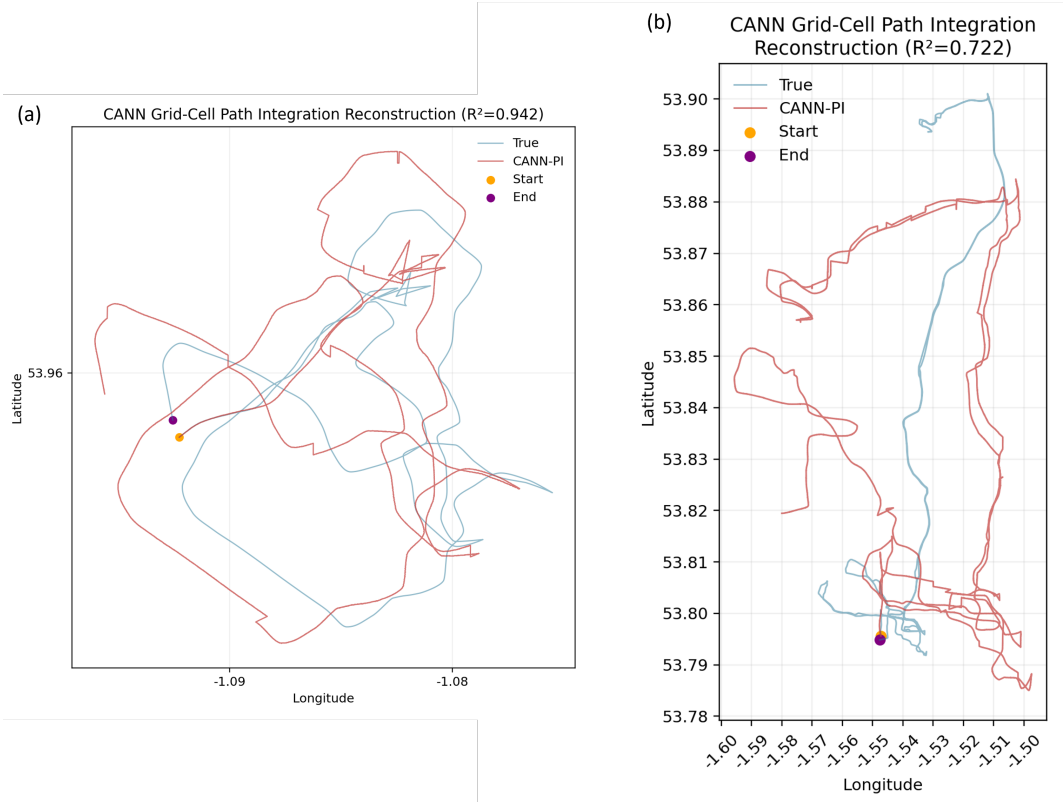


Figure 3 Path integration results based on logarithmically biased velocity signals in York (a) and Leeds (b).

Being driven by velocity signals is a defining feature of the grid-cell motion-tracking system. This velocity priority can explain several spatial cognitive phenomena. For example, when individuals travel between the same pair of places using different modes, such as walking one way and taking a bus the other, perceived speed biases may differ by direction. This can produce different distance estimates and generate the classic asymmetry of distance estimation in spatial cognition research [1]. Such asymmetry means that the same place may occupy different positions in cognitive space. Across a larger mobility network, non-integration may occur repeatedly. This provides a computational externalization and interpretation of the cognitive collage, characterized by locally near-Euclidean structure but non-uniform and fragmented global spatial relations.

Another issue concerns validation. Evaluating the alignment between "neurally estimated" cognitive distance and human judgments remains challenging. One feasible strategy is to follow multidimensional scaling (MDS) logic by asking survey participants to judge distances between pairs of places and comparing those judgments with the computational results. This could evaluate the logarithmic rule, but would not reveal how individual differences shape distorted cognitive mapping. A complementary strategy is to fit velocity perception parameters or functions to surveyed distance estimates. Future studies can investigate these questions in greater depth.

Undoubtedly, path integration is not the only component of cognitive mapping. In urban navigation, landmarks and other environmental features also play important roles. As travel time increases, pure path integration tends to generate larger errors in distance estimation,

much as walking in darkness accumulates positional uncertainty. Landmarks provide sensory information, especially visual input, that can correct current position estimates and help integrate the overall map. At the neural level, path integration and sensory correction correspond to the hippocampal-entorhinal interactive system represented by grid cells and place cells [20, 19]. A major task for future work is to understand how visible elements, including landmarks and boundaries, correct path-integration biases and enable more accurate cognitive-map modeling. This issue is particularly important in urban environments. How do urban-specific elements, such as transit lines, street network topology, and building distribution, affect cognitive mapping differently from a rat moving in an experimental arena? Contemporary computational neuroscience suggests that cognitive maps are built upon two major pillars: movement-derived reference frames and sensory feature anchors [9]. Modeling urban cognitive maps with such interdisciplinary tools may therefore renew how behavioral geography and geographic information science understand space.

References

1 Christian Bongiorno, Yulun Zhou, Marta Kryven, David Theurel, Alessandro Rizzo, Paolo Santi, Joshua Tenenbaum, and Carlo Ratti. Vector-based pedestrian navigation in cities. *Nature computational science*, 1(10):678–685, 2021. doi:10.1038/S43588-021-00130-Y.

2 Yoram Burak and Ila R Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009. doi:10.1371/JOURNAL.PCBI.1000291.

3 Elizabeth R Chrastil and William H Warren. From cognitive maps to cognitive graphs. *PloS one*, 9(11):e112544, 2014.

4 Helen Couclelis, Reginald G Golledge, Nathan Gale, and Waldo Tobler. Exploring the anchor-point hypothesis of spatial cognition. *Journal of environmental psychology*, 7(2):99–122, 1987.

5 Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017.

6 Tom M George, Mehul Rastogi, William de Cothi, Claudia Clopath, Kimberly Stachenfeld, and Caswell Barry. Ratinabox, a toolkit for modelling locomotion and neuronal activity in continuous environments. *Elife*, 13:e85274, 2024.

7 R G Golledge and L J Hubert. Some Comments on Non-Euclidean Mental Maps. *Environment and Planning A: Economy and Space*, 14(1):107–118, January 1982. doi:10.1068/a140107.

8 Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

9 Jeff Hawkins. *A thousand brains: A new theory of intelligence*. Basic books, 2021.

10 Sichao He, Aiersi Tuerhong, Shangjun She, Tianhao Chu, Yuling Wu, Junfeng Zuo, and Si Wu. Canns: Continuous attractor neural networks toolkit, February 2026. doi:10.5281/zenodo.18453893.

11 Mikail Khona and Ila R Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022.

12 Ed Manley, Gabriele Filomena, and Panos Mavros. A spatial model of cognitive distance in cities. *International Journal of Geographical Information Science*, 35(11):2316–2338, 2021. doi:10.1080/13658816.2021.1887488.

13 Meiyu Pan and Eve Isham. Temporal discounting in travel distance perception: Implications for transit design. In *International Conference on Transportation and Development 2025*, pages 830–841, 2025.

14 Luca Pappalardo, Ed Manley, Vedran Sekara, and Laura Alessandretti. Future directions in human mobility science. *Nature computational science*, 3(7):588–600, 2023. doi:10.1038/S43588-023-00469-4.

15 Waldo R Tobler. Geographic area and map projections. *Geographical review*, 53(1):59–78, 1963.

16 Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

17 Barbara Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *European conference on spatial information theory*, pages 14–24. Springer, 1993. doi:10.1007/3-540-57207-4_2.

18 William H. Warren. Non-Euclidean navigation. *Journal of Experimental Biology*, 222(Suppl_1):jeb187971, February 2019. doi:10.1242/jeb.187971.

19 James CR Whittington, David McCaffary, Jacob JW Bakermans, and Timothy EJ Behrens. How to build a cognitive map. *Nature neuroscience*, 25(10):1257–1272, 2022.

20 James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

21 Yanling Zuo, Ruoyu Chen, and Jia Zhou. Impact of velocity changes of surrounding vehicles on the eye movement, subjective cognition, and driving behaviour of novice drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 114:423–440, 2025.

Towards a Validation Taxonomy for Agent-Based Models of Pedestrian Movement in Cities

Marcin Wozniak*
Adam Mickiewicz University in Poznan, Poland

Gabriele Filomena*
Department of Geography and Planning, University of Liverpool, UK

Abstract

Agent-Based Models (ABMs) are increasingly used to simulate and analyse pedestrian movement in cities and to support urban planning and policy decisions. However, their credibility depends critically on rigorous empirical validation, for which no standardised framework currently exists. This paper addresses this challenge by providing a structured synthesis of existing validation strategies, tailored to ABMs of pedestrian movement in cities, focusing on the comparison of simulated and reference behaviours. First, it reviews and systematises operational validation approaches spanning three model output levels: micro, meso, and macro. Then, it develops a validation taxonomy that links specific model outputs – individual behavioural trajectories, segment-level flows, and aggregate system indicators – to data sources, potential validation methods, and appropriate behavioural claims. Finally, it demonstrates the implementation of these approaches using illustrative examples. Here, it is argued that not only do different model outputs require distinct data sources and methodological strategies, but the generalisations a validation exercise can support are bounded by what is validated and how. The paper advances a practical validation taxonomy that, whilst setting standards in the discipline, can be flexibly adapted to diverse modelling contexts. In doing so, it supports the development of more robust, transparent, and policy-relevant agent-based models of pedestrian movement in cities.

2012 ACM Subject Classification Computing methodologies → Multi-agent systems; Computing methodologies → Intelligent agents; Computing methodologies → Model verification and validation; Computing methodologies → Simulation evaluation

Keywords and phrases agent-based modelling, pedestrian movement, route choice behaviour, validation, cities

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.35

Category Short Paper

1 Introduction

Agent-based modelling is a computational paradigm that explores the behaviour of autonomous entities such as humans, social organisations, and animals in their environment. This framework is widely adopted to simulate, explore, and understand pedestrian dynamics in urban space. Agent-Based Models (ABMs) of pedestrian movement in cities have been developed at a variety of geographical scales, ranging from large urban areas to districts and neighbourhoods [4, 15]. Varying levels of behavioural complexity have been incorporated in pedestrian agents' route choice behaviour – that is, how agents formulate routes between different locations in the simulated environment – with many models resorting to shortest-path-based approaches.

Although recent modelling efforts have substantially increased the realism and sophistication of ABMs of pedestrian movement in cities, empirical validation has not progressed at the same pace [14, 13]. This gap matters: validation "relates to the extent that the model

* These authors contributed equally to this work.

© Marcin Wozniak and Gabriele Filomena;
licensed under Creative Commons License CC-BY 4.0

adequately represents the system being modelled" [3, p. 419], or, more operationally, whether "the model's accuracy is within its acceptable range of accuracy" [17, p. 12]. Several challenges contribute to this gap, including: i) limited availability of open, high-quality pedestrian data, and persistent privacy constraints [11]; ii) the cost and technical complexity involved in collecting and handling empirical validation data [7]; iii) the absence of universally accepted standards or one-size-fits-all approaches to validation in agent-based modelling [2]. Concerns about the reliability [4] of model outputs not only diminish the methodological relevance of the work, but also limit the ability of these models to shed light on principles of pedestrian movement and behaviour as a theoretical contribution, and advance policy-relevant insights or planning solutions.

ABMs of pedestrian movement in cities are used for different purposes: to explain and understand existing movement dynamics, test potential planning or design interventions, and estimate pedestrian movement under alternative system conditions. These purposes do not require the same form of validation. A model intended to reproduce aggregate pedestrian volumes, for instance, should not necessarily be evaluated in the same way as a model intended to represent individual route choice, behavioural adaptation, or localised responses to environmental conditions. Validation therefore needs to be aligned with the model output and with the claims being made.

This paper makes three contributions. First, it develops a structured taxonomy linking micro-, meso-, and macro-level model scopes and outputs to appropriate validation strategies. Second, it provides practical, step-by-step examples demonstrating how particular data sources can be leveraged for empirical validation through different operationalisation. Finally, it systematises existing validation strategies in the context of ABMs of pedestrian movement in cities and proposes a framework for linking model outputs, empirical data sources, validation strategies, and the behavioural claims they can support.

2 Operational validation in ABMs of pedestrian movement in cities

In this work, following [17], validation is approached as a problem of *operational validity*, that is, whether the output patterns of a simulation model have a satisfactory level of accuracy for the model's intended purpose and domain of applicability. The central argument is that different model outputs support different behavioural claims and therefore require different validation strategies. In ABMs of pedestrian movement in cities, the same model may generate several forms of output behaviour, including individual trajectories, route choices, walking times, segment-level pedestrian flows, spatial distributions of agents, temporal activity profiles, and system-level relationships.

In this direction, we distinguish between two operational validation procedures: exploring model behaviour and comparing output behaviours. On the one hand, *exploring model behaviour* refers to the systematic examination of the outputs produced by the ABM under different parameter values, behavioural assumptions, spatial representations, and experimental conditions. This may include sensitivity analysis, scenario inspection, and assessment of whether simulated patterns respond plausibly to changes in the model specification. This procedure is especially important when direct empirical observations are unavailable, incomplete, or not directly comparable with model outputs. On the other hand, *comparing output behaviours* refers to the direct assessment of simulated outputs against empirical observations, outputs from other models, or theoretical benchmarks. In observable systems, this comparison can be made against empirical pedestrian data, such as GPS traces, video-derived trajectories, pedestrian counts, aggregated mobile-device data, or platform-derived activity indicators.

This logic underpins the taxonomy used here. We focus on comparing output behaviours, organised across three output levels: micro-level validation concerns individual agent properties and emerging behaviours, targeting *behavioural consistency*; meso-level validation concerns collective movement flows across space and time, targeting *spatial and temporal consistency*; and macro-level validation concerns aggregated system-level patterns and model-wide distributions, targeting *aggregated logical consistency* or *theoretical plausibility*. These validation levels are analytically related but not interchangeable. Micro-level outputs may aggregate into meso-level flows, and meso-level patterns may contribute to macro-level system behaviour. However, each level captures a different abstraction of pedestrian movement and supports different claims. Validation should therefore emerge from the model purpose and output structure, rather than be carried out as a generic post-hoc exercise.

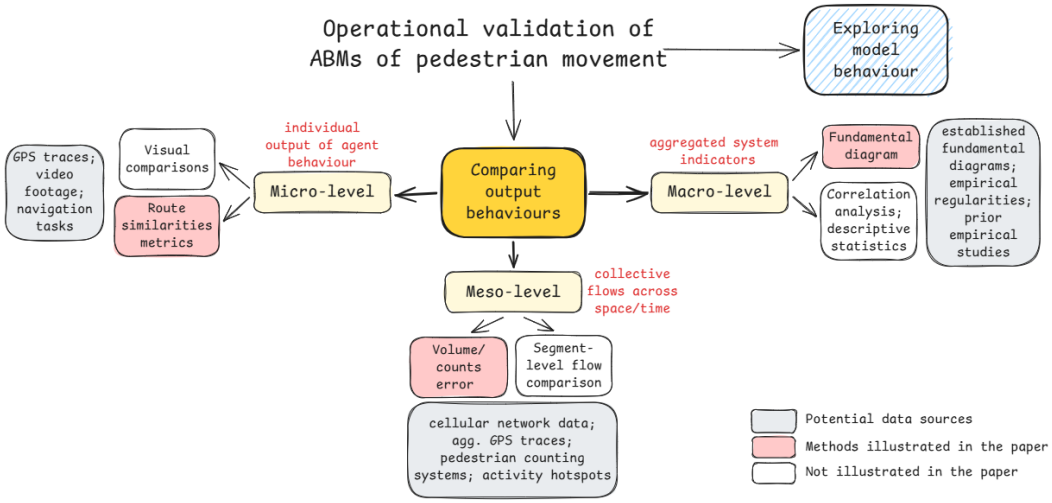


Figure 1 Approaches to *operational validation* of ABMs of pedestrian movement in cities.

2.1 Illustrative ABM case studies

To schematise potential validation strategies, we draw on two illustrative agent-based modelling case studies based on different modelling principles. The first ABM, hereafter referred to as *Case Study I*, simulates pedestrian movement in the Old Market Square in Poznań, Poland[1] [20]. In this model, agents enter and exit the simulated urban space through gateways linked to transport nodes. They then navigate towards randomly assigned Points of Interest (POIs) distributed across the square using a gradient-based movement mechanism, before returning to their entry gateway. As agents follow a gradient-based mechanism, this model is oriented towards explaining collective and system-level movement dynamics, and is used here for meso- and macro-level validation.

The second ABM, referred to as *Case Study II*, simulates pedestrian movement in the historical centre of Poznań. In this case, agents compute routes on a primal graph representation of the street network and travel from home locations to randomly selected POIs [21]. Here, as agents compute explicit routes, the model is oriented towards representing individual route choice, and is used for micro-level validation against observed trajectories.

[1] Poznań is a city located in west-central Poland, with a population of approximately 550,000 inhabitants; Old Market Square is the main square in the city centre.

2.2 Micro-level validation: Individual agent properties and emerging behaviours

Micro-level patterns refer to variables that capture individual behaviours. These include agent states (e.g. position, heading, speed, acceleration patterns, pedestrian interactions), as well as direct manifestations of behaviour, such as individual movement trajectories or distance travelled. Here, the validation exercise involves using high-resolution spatio-temporal data to verify the accuracy of simulated pedestrian behaviour against real-world movement. Research questions at this level of abstraction investigate, for example, the extent to which the agent routes match those observed in empirical data. In some cases, when the ABM explores more granular behaviour, these questions could even refer to locomotion details, avoidance behaviours, and so forth.

To illustrate this, we compared GPS traces recorded during a navigation task, in which 18 participants completed routes between fixed pairs of Origin and Destination (OD) points (36 routes in total) [22], with routes generated in *Case Study II* for the same OD pairs. Figure 2 displays both the agent routes generated by the ABM and the observed GPS traces.

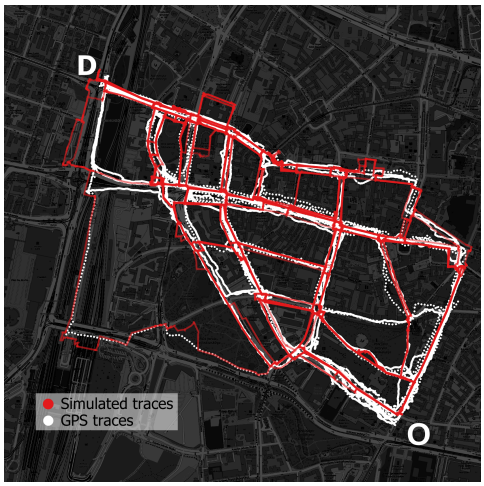


Figure 2 Simulated routes from *Case Study II* and the corresponding observed GPS traces.

While differences between individual trajectories and simulated outputs might be identified visually, several techniques can be used to quantify the divergence between simulated and observed routes, including Dynamic Time Warping, Fréchet distance, entropy-based measures, and directional change analysis [18]. In *Case Study II*, we adopted *point-to-point Euclidean error* as the accuracy metric. Specifically, the mean *route discrepancy* was computed as the average point-to-point distance per route, then averaged across the 36 routes, with smaller values indicating greater similarity between simulated agent behaviour and observed pedestrian movement. Furthermore, we compared measures of central tendency for route length and walking time across simulated outputs and empirical GPS data (Table 1).

Table 1 Properties of the simulated routes from *Case Study II* and the corresponding observed GPS traces (means across 36 routes from 18 participants).

Route Length (m) observed	Route Length (m) simulated	Route Discrepancy (m) mean / median	Walking Time (min.) observed	Walking Time (min.) simulated
2030	2014	5.2 / 3.52	20.04	20.58

This approach heavily depends on the availability of high-resolution trajectory data and is sensitive to sampling bias, positional error, map-matching choices, and the representativeness of the observed routes. Despite these challenges, micro-level validation remains one of the most behaviourally detailed approaches available for empirically validating ABMs of pedestrian movement. Crucially, accuracy at this level does not by itself support claims about collective flows or system-level dynamics.

2.3 Meso-level validation: Collective movement flows across space and time

Meso-level outputs refer to the spatial and temporal distribution of pedestrian agents. These outputs may include agent counts on street-network segments, pedestrian volumes at representative locations, such as city squares, green areas, or bus stops, or counts aggregated to spatial units such as census areas or hexagonal cells. When temporal dynamics are relevant, these counts can also be aggregated by time interval, for example to compare pedestrian activity across different times of day.

The most direct validation strategy at this level is to compare simulated outputs with observed pedestrian counts. Municipal pedestrian counting systems[2] provide one such source, as they rely on counting devices distributed across parts of the street network. These datasets can offer high temporal granularity, although their suitability for validation depends on the number and spatial coverage of the sensors. A second option is to make use of datasets that describe human activity hotspots [12, 9]. Such data indirectly capture how pedestrian presence varies across locations and time, and may be derived from aggregated GPS traces, cellular network records, or other crowdsourced signals. An example is Google's Popular Times service, which provides anonymised estimates of place-based busyness throughout the day and week. These datasets may represent pedestrian activity either as relative intensity indices or as counts associated with specific points of interest or street segments. They can be accessed through APIs or open data portals and may provide broad spatial and temporal coverage across urban environments. Images obtained from security cameras, video sensors, or scene-analysis methods can also be used to infer pedestrian presence and constitute an alternative source of meso-level validation data. Recent work, for example, has estimated street segment pedestrian volumes from Street View images of Tianjin (China) by leveraging machine-learning pedestrian detection techniques [1].

Validation research questions at this level may therefore investigate whether the model can reproduce observed spatial and temporal patterns of pedestrian presence across the simulated space. The geographical representation may be the street network itself, a set of relevant locations, or aggregated spatial units, at different scales of analysis. For illustrative purposes, Google Popular Times data for Old Market Square in Poznań were used to derive proportional activity levels around places and, therefore, as a proxy for pedestrian volumes during a Sunday. The upper panel of Fig. 3 presents the aggregated pedestrian volumes generated by *Case Study I* and compares them with pedestrian traffic derived from Google Popular Times[3]. The lower panel displays the error between simulated and empirical values.

The scaled series can then be compared using standard error metrics. In this illustrative case, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were 29.86, 20.38, and 19.57, respectively.

[2] Sensor-based pedestrian count data are available for e.g. Melbourne (`https://www.pedestrian.melbourne.vic.gov.au/`), New York (`https://www.nyc.gov/html/dot/html/about/datafeeds.shtml`), or the City of Casey (`https://data.casey.vic.gov.au/`).

[3] Pedestrian numbers were scaled to a 0–1000 interval to make the two series comparable.

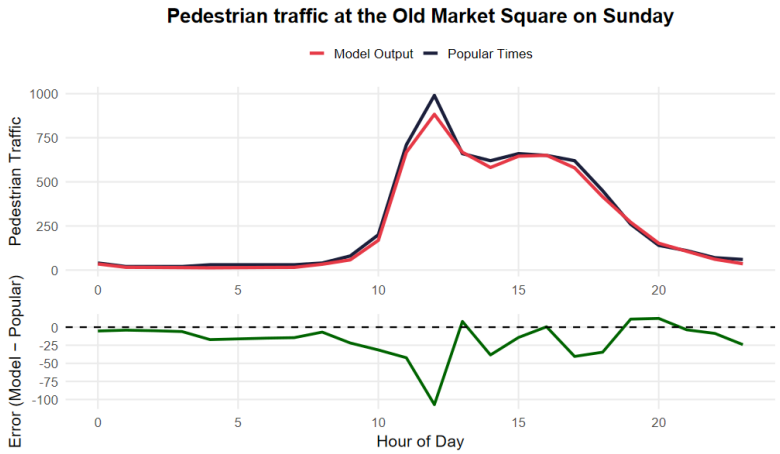


Figure 3 Popular Times-derived pedestrian activity data and the outputs of *Case Study I*.

Finally, existing work also highlights the potential value of adopting complementary techniques for validating meso-level outputs. Pouke and colleagues [16], for example, used mobility traces from a municipal Wi-Fi network to construct a time-varying OD matrix, which was then treated as a ground-truth representation of pedestrian movement across locations and time periods. Simulated pedestrian footprints were compared against this reference to identify where the model over- or under-estimated pedestrian presence at different times of day. Similarly, Filomena and Verstegen [5] evaluated simulated pedestrian volumes at the street segment level by deriving observed segment counts from a set of overlaid GPS trajectories. Their validation combined a spatial diagnosis of over- and underestimation across individual street segments with a network-level error assessment based on RMSE.

The scope of the claims that one can advance at this level is bounded: reproducing aggregate spatial and temporal patterns of presence supports deriving insights about collective pedestrian activity, but not about individual behavioural mechanisms generating those patterns.

2.4 Macro-level validation: System-level behavioural patterns

Macro-level and low-granularity aggregated outputs capture system-level characteristics, such as pedestrian flow rates, total volumes, average route length, density–speed relationships, and other aggregate regularities, and are commonly used to assess the overall logic of the simulated system. These outputs, rather than being validated directly in terms of fine-grained empirical accuracy, can be scrutinised to provide a general evaluation of the model.

Along these lines, for example, the *fundamental diagram* of pedestrian flows describes the relationship between pedestrian density, walking speed, and flow [19]. It typically shows that, as density increases, walking speed decreases, while flow – defined as $density \times speed$ – initially increases, reaches a maximum, and then declines under congested conditions [8]. In this context, the fundamental diagram serves as a theory-driven benchmark rather than a strictly location-specific empirical reference. The fundamental diagrams derived from the outputs of *Case Study I* are presented in Fig. 4, while the corresponding regression models are reported in Table 2

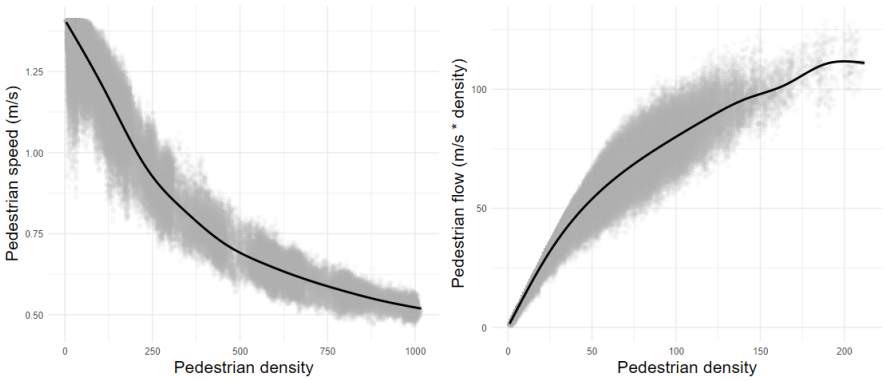


Figure 4 Fundamental diagrams of the simulated pedestrian flows in *Case Study I*.

Table 2 Coefficients of speed–density (regression model I) and flow–density (regression model II) relationships for the outputs of *Case Study I*.

Regression model	**I**	**II**
const	1.34^{***}	6.54^{***}
	(0.000)	(0.000)
speed	-1.10^{***}	–
	(0.000)	
flow	–	0.79^{***}
		(0.000)
R^2	0.92	0.93
F statistics p-value	0.000	0.000

Note: p-values are reported in brackets.

This sort of outputs can be examined using either location-specific data or studies that have estimated fundamental diagrams in different urban contexts. For example, empirical studies consistently confirm a negative relationship between pedestrian speed and density, with reported coefficients of determination (R^2) ranging from 0.78 to 0.99 [10]. However, estimated regression parameters vary across studies, with intercepts typically ranging from 1.0 to 2.0 and density coefficients from -0.7 to -1.3 [6].

Along similar lines, modellers have compared aggregated model outputs with empirical evidence reported in the literature with regards to, for instance, pedestrian preferences or behavioural characteristics. Although these assessments look at micro-level behaviour, they typically evaluate it through aggregated properties of agent routes, such as percentage deviation from the shortest path, and compare these outputs with empirical research on the determinants of route-choice behaviour. This type of assessment, when used to identify which ABM scenario or model configuration performs better against observational data, evaluates the behavioural plausibility of the model. One could argue that macro-level validation supports explanatory plausibility rather than direct predictive accuracy, as it assesses whether simulated outputs reproduce theoretically and empirically established patterns of pedestrian behaviour.

Unlike other approaches, this form of validation does not necessarily require external datasets specific to the case study. Instead, it relies on well-documented empirical regularities in pedestrian research. Nevertheless, several limitations should be noted. First, assessments based on the fundamental diagram refer to generalised empirical regularities rather than

context-specific measurements. Second, pedestrian behaviour may deviate substantially from theoretically derived relationships because of individual heterogeneity and specific environmental or spatial configurations. Third, this type of validation does not directly evaluate route-choice behaviour, pedestrian spatial distribution, or temporal fluctuation patterns. As such, macro-level validation strategies, including those based on the fundamental diagram, should be understood as complementary rather than stand-alone validation approaches.

3 Synthesis: Towards a validation taxonomy

The examples above show that validating ABMs of pedestrian movement in cities is not a single methodological operation, but a structured process of aligning model scope, outputs, reference evidence, validation methods, and modelling claims. Table 3 presents a preliminary validation taxonomy that links levels of model output to corresponding data sources, methodological approaches, and the types of claims and insights the ABM can support. The taxonomy concerns the comparison of output behaviours, and it assumes that reference evidence is available; when direct comparison is not possible, the exploration of model behaviour (see above) remains the main alternative. The taxonomy is not intended to prescribe a universal metric or fixed validation threshold. Rather, it offers a reasoning structure for making validation choices explicit and comparable across ABMs of pedestrian movement.

Table 3 A preliminary validation taxonomy for pedestrian ABMs, linking output levels and model outputs to validation targets, potential data sources, methodological approaches, and the types of claims each can support.

Output level	Model outputs	Validation target	Reference evidence	Validation methods	Claims supported
Micro-level	Individual routes, trajectories, walking times, travelled distance	Behavioural consistency of individual agents	GPS traces, video tracking, lab-based navigation data, observed route choices, etc.	Route overlap, point-to-point error, Fréchet distance, Hausdorff distance, Dynamic Time Warping, entropy-based measures, directional change analysis, travel-time comparison, etc.	Claims about individual route choice, wayfinding, behavioural adaptation, decision rules, and agent-level movement realism
Meso-level	Segment-level flows, agent volumes, spatial concentrations of agents, temporal activity profiles	Spatial and temporal consistency of collective movement	Pedestrian counters, aggregated GPS traces, mobile-phone data, sensor networks, platform-derived activity indicators (e.g. Popular Times), image-derived counts, etc.	RMSE, MAE, MAPE, GEH statistic, temporal profile comparison, density surface comparison, segment-level flow comparison, hotspot comparison, spatial correlation analysis, etc.	Claims about collective pedestrian presence, public-space usage, temporal fluctuation, activity intensity, and local flow patterns
Macro-level	Aggregate system indicators, density–speed and density–flow relations, average route length, network-level measures	Aggregated logical consistency or theoretical plausibility of system dynamics	Published fundamental diagrams, established empirical regularities, outputs from other models, prior empirical studies, etc.	Distributional similarity, comparison of estimated coefficients with empirical ranges, comparison with theoretical expectations or benchmark simulation models, etc.	Claims about internal consistency, emergent patterns, theoretical plausibility, and congestion dynamics

The taxonomy addresses a common ambiguity in agent-based modelling for pedestrian movement simulation: treating visual plausibility, aggregate fit, behavioural realism, and theoretical consistency as equivalent forms of validation. From a spatial information perspective, the taxonomy highlights that validation is inseparable from spatial abstraction. ABMs of pedestrian movement in cities go beyond simulating pedestrian trajectories; they encode assumptions about space, scale, topology, accessibility, interaction, and decision-making. The output level at which a model is validated should also be aligned with its purpose: models used for prediction require stronger empirical comparison at the relevant output level, whereas models intended for explanation or scenario testing may resort to behavioural plausibility and system level consistency. This proposed taxonomy should therefore be understood as a

flexible framework for designing and reporting validation exercises. At minimum, a validation claim should specify: i) which model output is being assessed; ii) which reference evidence is used; iii) which metric, diagnostic, or benchmark is applied; and iv) what claims the validation result supports. Making these elements explicit would improve transparency and reproducibility across agent-based modelling of pedestrian movement.

References

1 Long Chen, Yi Lu, Qiang Sheng, Yu Ye, Ruoyu Wang, and Ye Liu. Estimating pedestrian volume using Street View images: A large-scale validation test. *Computers, Environment and Urban Systems*, 81:101481, 2020. doi:10.1016/j.compenvurbsys.2020.101481.

2 Andrew Collins, Matthew Koehler, and Christopher Lynch. Methods That Support the Validation of Agent-Based Models: An Overview and Discussion. *Journal of Artificial Societies and Social Simulation*, 27(1):11, 2024. doi:10.18564/jasss.5258.

3 Andrew T. Crooks, Christian Castle, and Michael Batty. Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems*, 32(6):417–430, 2008. doi:10.1016/j.compenvurbsys.2008.09.004.

4 Andrew T. Crooks, Arie Croitoru, Xu Lu, Sarah Wise, John M. Irvine, and Anthony Stefanidis. Walk This Way: Improving Pedestrian Agent-Based Models through Scene Activity Analysis. *ISPRS International Journal of Geo-Information*, 4(3):1627–1656, 2015. doi:10.3390/ijgi4031627.

5 Gabriele Filomena and Judith A. Verstegen. Modelling the effect of landmarks on pedestrian dynamics in urban environments. *Computers, Environment and Urban Systems*, 86:101573, March 2021. doi:10.1016/j.compenvurbsys.2020.101573.

6 Maria Giannoulaki and Zoi Christoforou. Pedestrian Walking Speed Analysis: A Systematic Review. *Sustainability*, 16(11):4813, 2024. doi:10.3390/su16114813.

7 Zhou He, Qi Song, Junchao Lian, and Yiming Liu. Towards Standardizing Validation Practices in Agent-Based Modeling: A Hierarchical ABM Validation Framework. *ACM Transactions on Modeling and Computer Simulation*, 36(1):1–25, 2026. doi:10.1145/3769857.

8 Serge P. Hoogendoorn, Winnie Daamen, Victor L. Knoop, Jeroen Steenbakkers, and Majid Sarvi. Macroscopic Fundamental Diagram for pedestrian networks: theory and applications. *Transportation Research Procedia*, 23:480–496, 2017. doi:10.1016/j.trpro.2017.05.027.

9 Tao Jia and Zheng Ji. Understanding the Functionality of Human Activity Hotspots from Their Scaling Pattern Using Trajectory Data. *ISPRS International Journal of Geo-Information*, 6(11):341, 2017. doi:10.3390/ijgi6110341.

10 Kotkar Kishor Laxman, Rajat Rastogi, and Satish Chandra. Pedestrian Flow Characteristics in Mixed Traffic Conditions. *Journal of Urban Planning and Development*, 136(1):23–33, 2010. doi:10.1061/(ASCE)0733-9488(2010)136:1(23).

11 Shiqin Liu, Carl Higgs, Jonathan Arundel, Geoff Boeing, Nicholas Cerdera, David Moctezuma, Ester Cerin, Deepti Adlakha, Melanie Lowe, and Billie Giles-Corti. A Generalized Framework for Measuring Pedestrian Accessibility around the World Using Open Data. *Geographical Analysis*, 54(3):559–582, 2022. doi:10.1111/gean.12290.

12 Roberto Ponce Lopez and Joseph Ferreira. Identifying spatio-temporal hotspots of human activity that are popular non-work destinations. *Environment and Planning B: Urban Analytics and City Science*, 48(3):433–448, 2021. doi:10.1177/2399808320970209.

13 Nick Malleson, Mark Birkin, Daniel Birks, Jiaqi Ge, Alison Heppenstall, Ed Manley, Josie McCulloch, and Patricia Ternes. Agent-based modelling for Urban Analytics: State of the art and challenges. *AI Communications*, 35(4):393–406, 2022. doi:10.3233/AIC-220114.

14 Steven Manson, Li An, Keith C. Clarke, Alison Heppenstall, Jennifer Koch, Brittany Krzyzanowski, Fraser Morgan, David O'Sullivan, Bryan C. Runck, Eric Shook, and Leigh Tesfatsion. Methodological Issues of Spatial Agent-Based Models. *Journal of Artificial Societies and Social Simulation*, 23(1):3, 2020. doi:10.18564/jasss.4174.

15 Itzhak Omer and Nir Kaplan. Structural properties of the angular and metric street network's centralities and their implications for movement flows. *Environment and Planning B: Urban Analytics and City Science*, 46(6):1182–1200, 2019. doi:10.1177/2399808318760571.

16 Matti Pouke, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. Practical simulation of virtual crowds using points of interest. *Computers, Environment and Urban Systems*, 57:118–129, 2016. doi:10.1016/j.compenvurbsys.2016.02.004.

17 R G Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1):12–24, 2013. doi:10.1057/jos.2012.20.

18 Yaguang Tao, Alan Both, Rodrigo I. Silveira, Kevin Buchin, Stef Sijben, Ross S. Purves, Patrick Laube, Dongliang Peng, Kevin Toohey, and Matt Duckham. A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing*, 58(5):643–669, 2021. doi:10.1080/15481603.2021.1908927.

19 Lakshmi Devi Vanumu, K. Ramachandra Rao, and Geetam Tiwari. Fundamental diagrams of pedestrian flow characteristics: A review. *European Transport Research Review*, 9(4):49, 2017. doi:10.1007/s12544-017-0264-6.

20 Marcin Wozniak. From dawn to dusk: daily fluctuations in pedestrian traffic in the city center. *SIMULATION*, 100(3):245–263, 2024. doi:10.1177/00375497231212543.

21 Marcin Wozniak and Gabriele Filomena. Beyond urban form: Introducing dynamic factors in agent-based simulations of pedestrian movement. *Computers, Environment and Urban Systems*, 128:102439, 2026. doi:10.1016/j.compenvurbsys.2026.102439.

22 Marcin Wozniak, Gabriele Filomena, and Adam Wronkowski. What's your type? A taxonomy of pedestrian route choice behaviour in cities. *Transportation Research Part F: Traffic Psychology and Behaviour*, 109:1257–1274, 2025. doi:10.1016/j.trf.2025.01.012.

From Sequential Decisions to Concurrent Processes: Advancing Human Navigation Models for Simulation

Sabine Timpf
Geoinformatics Group, Institute of Geography, University of Augsburg, Germany

Franziska Klügl
School of Science and Technology, University of Örebro, Sweden

Abstract

Computational models of human navigation have largely followed sequential decision paradigms, treating planning, perception, and movement as discrete models. In contrast, evidence from spatial cognition and neuroscience indicates that navigation arises from multiple interacting processes operating concurrently. Building on a previously proposed architecture, this paper advances a modelling framework that explicitly represents such concurrency and process interaction, and examines its implications for agent-based simulations of spatial decision-making in human navigation.

2012 ACM Subject Classification Computing methodologies → Model development and analysis; Computing methodologies → Intelligent agents; Computing methodologies → Multi-agent systems

Keywords and phrases agent-based simulation, modelling human navigation, computational process modelling

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.36

Category Short Paper

Supplementary Material
Software (Source Code): `https://github.com/GeosavvyAgents/concurrentNavigationProcesses` archived at `swh:1:dir:8a5ea68e524352a605d85e33b2f7132b9c54bfce`

Funding *Sabine Timpf*: I gratefully acknowledge support by University of Örebro for my research stay in March 2026.

1 Human Navigation encompasses concurrent processes

Navigation in spatial environments cannot be reduced to a single decision-making process. Instead, navigation behaviour emerges from the interaction of multiple cognitive and sensorimotor processes, including perception, route planning, orientation, path following, and movement [17, 5, 24]. These processes operate at different spatial and temporal scales and continuously influence one another during navigation [1]. In addition, they are based on different representations of the spatial context at multiple levels of abstraction [27, 26].

For agent-based simulation, this implies that realistic models must account for the simultaneous and ongoing interplay between higher-level wayfinding and lower-level locomotion, rather than treating these processes as independent or sequential stages of the overall navigation process. However, most existing agent-based approaches to simulating navigation rely on architectures derived from the sense–plan–act paradigm [23]. These architectures typically impose a sequential control structure or introduce hierarchical dependencies between behavioural levels. As a result, they struggle to represent navigation as a dynamic, situated activity in which processes such as movement, re-orientation, and route-(re)planning occur in parallel and interact continuously. While hybrid models combine planning and reactive movement, they often do so within tightly coupled or centrally controlled frameworks. What

© Sabine Timpf and Franziska Klügl;
licensed under Creative Commons License CC-BY 4.0

remains largely unsupported is an explicit representation of navigation as a set of concurrent processes across scales that can operate independently while still interacting through the spatial context.

In this paper, we report on our progress in representing and simulating human navigation as interacting concurrent processes (routing, path following, orienting, moving) using shared states (places, paths, route) as interleaving principle. By shifting the focus from isolated decisions to interacting processes, we argue for a new generation of agent-based models for spatial behaviour that better reflect the complexity of human navigation. We expect to contribute to Spatial Information theory by introducing a conceptual and methodological framework for modelling human navigation as a set of explicitly represented, interacting concurrent processes, and demonstrating its implications for agent-based simulations of spatial decision-making.

The paper is structured as follows: After a short review of computational models of human navigation in section 2, we discuss the approaches for modelling concurrent processes and their interactions (section 3), and present our approach for modelling and simulating human navigation as loosely coupled processes coordinated through informational states (section 4). We then discuss the current state of the research (section 5) and present our (preliminary) conclusions as well as some future work in section 6.

2 Modelling Human Navigation

Research on human navigation spans cognitive science, geography, robotics, and simulation, yet remains fragmented with respect to computational process modelling. Reviews of spatial memory (e.g., [16]) emphasize internal representations, while pedestrian and crowd models (e.g., [12, 18]) focus on movement dynamics and local interactions. Behavioural accounts [8] describe navigation as a multi-stage activity, but are rarely formalized computationally. In contrast, robotics approaches (e.g., [25]) provide explicit computational frameworks, yet typically rely on sequential sense–plan–act architectures. Agent-based models integrate cognitive and environmental factors (e.g., [3, 6]), but generally adopt modular or staged designs. Across these strands, navigation is decomposed into components, but not modelled as a system of interacting, concurrent processes.

While many computational models acknowledge that behaviour emerges from multiple interacting processes, they rarely treat these processes as explicitly concurrent at the architectural level. Early cognitive models such as NAVIGATOR [9] and PLAN [2] decompose navigation into functional components, but implement them largely as sequential reasoning steps without continuous interaction with locomotion. More recent pedestrian simulations – such as those by Kneidl et al. [14], Gath-Morad et al. [7], and Kielar et al. [13] – introduce multi-layered structures that separate high-level routing from low-level movement. These approaches implicitly recognize concurrency, for example by combining path planning with reactive collision avoidance, yet interactions remain tightly coupled or hierarchically controlled.

The work of Torrens [28] is particularly notable in this context. His agent-based framework integrates path planning, environmental perception, and locomotion within a unified system, enabling dynamic adaptation to context and other agents. However, even here, concurrency is realized through tightly integrated modules within a single decision framework rather than as independent, asynchronously operating processes.

In contrast, the approach proposed in this paper treats concurrency as a first-class modelling principle: navigation is represented as a set of loosely coupled processes operating in parallel and interacting through shared state, rather than through centralized control or sequential coordination. This distinction enables a more transparent and flexible representation of the interdependencies that characterize human navigation.

The spatial scale addressed in this paper is the local scale of individual navigation. The framework is concerned with agents moving through environments such as buildings, campuses, parks, street networks, or neighbourhood-scale settings, where navigation involves route following, orientation, perception of local cues, and possible adjustment of movement or route-related decisions. It is not primarily aimed at regional transport modelling, large-scale activity scheduling, or aggregate travel-demand forecasting, although such models could provide higher-level constraints or destinations.

3 Modelling concurrent processes and their interactions

Concurrent processes in core computer science are commonly modelled using formalisms from distributed systems [29]. There, a basic distinction is made between message-based and shared-memory-based paradigms. In message-based modelling languages, such as Communicating Sequential Processes or the π-calculus, processes exchange information through explicit communication, whereas shared-memory approaches coordinate processes through common variables. Communicating automata similarly model interaction by showing how messages or events trigger transitions and produce further events. For modelling concurrent processes within one agent's decision-making, however, the key issue is not the communication mechanism itself, but the dependencies between processes: information needs, enabling conditions, and required updates. In a message-based approach, process A sends information to process B; in a shared-memory approach, process A writes information to a variable that process B monitors. In both cases, the relevant contribution is to make inter-process information dependencies explicit.

Thus, for the level of abstraction that is relevant here, the shared memory-based approach appears to be more appropriate. Clearly, we ignore issues related to performance, robustness or the complexity of synchronizing shared access avoiding race conditions. At this level of abstraction, the major languages to capture process behaviour and interaction mostly associated with state changes with explicit conditions (events) are Petri-Nets and more basic statechart-like languages that allow to formulate more condition-driven coordination between concurrent processes.

Petri nets [22], in particular in their condition–event interpretation, model system behaviour in terms of conditions represented by places and events represented by transitions, with tokens placed in condition nodes indicating the current state of the system. A transition is enabled when all required input conditions are satisfied (have a token), making dependencies between process steps explicit. While this provides a precise representation of causal relationships, it is not always easy to distinguish clearly how different concurrent processes develop due to the tight coupling between overall system state representation and conditions for process progression.

Another language that occurs in the context of modelling concurrent processes – more for verification – is TLA+[15]. It is a state-based specification language based on temporal logic that characterizes overall system behaviour in terms of transitions between states given logical constraints. Transitions happen based on actions from different concurrent processes. While it can be used to model concurrent processes, it does not provide an explicit notion of process

structure, instead it captures concurrency implicitly through nondeterministic interleaving of enabled actions. Yet, we need explicit representation of interdependent processes and their progression, not just constraints on global behaviour.

Statecharts extend classical automata-based models by introducing hierarchical structuring, explicit support for formulating multiple loosely coupled processes that advance concurrently combined with condition-based transitions between different (sub-)process states. Those transitions can be guarded by conditions using the overall agent's informational state, enabling the explicit specification of dependencies between processes. Variants of statechart-based modeling are widely used in software engineering practice including the original formulation by D. Harel [10] to UML state machines [19].

A related question refers to how the agent's informational state – representing the shared memory for the concurrent processes – can be represented. In a simple object-oriented agent implementation, this may be represented as agent-level variables that are read and updated by the different concurrent internal processes. More elaborate agent architectures may use a belief base, in which information about the agent, its environment, and relevant contextual knowledge can be represented explicitly, often using some logic-based language as in BDI-type of architectures [21]. A further alternative is a blackboard architecture, where the shared state is more like a common problem-solving workspace [11]. Tuple-space models, such as Linda, provide a related coordination mechanism in which processes interact through a shared repository of structured data items [20]. Finally, publish–subscribe infrastructures (as in ROS) can be used when processes should remain more decoupled: rather than directly accessing shared variables, processes publish updates and subscribe to information types or events relevant to their own progression. These alternatives differ mainly in how explicit, structured, and decoupled the shared informational state is. Agent-level variables are simple and efficient but provide little semantic structure. Belief bases and blackboard systems make the informational content more explicit and extensible. Tuple spaces and publish–subscribe mechanisms further decouple information producers and consumers, but may introduce additional architectural complexity.

4 Proposed Approach for Simulating Concurrent Processes for Human Navigation

Our approach is conceptually related to blackboard architectures [11], in which multiple independent processes interact through a shared state without a fixed control sequence. While such architectures are no longer widely used as a software engineering paradigm, their core principle of state-mediated, opportunistic interaction remains influential in contemporary models of cognition and distributed systems.

We conceptualize the core navigation of pedestrians as consisting of at least four processes, i.e., routing, path following, orienting (also called self-localization), and walking, as originally suggested by Downs and Stea [4]. Figure 1 shows the conceptual model underlying each of the core navigation processes. It demonstrates how routing, path following, orienting, and walking each carry out their own reasoning, from surface-level motion, called walking, to abstract graph routing. This also implies different spatial representations for each reasoning process, shown on the right side of figure 1.

The orienting process stores orientation information in a memory structure, which is also used to compare locations, and checks if the goal place is reached. In the walking process, the agent moves from a starting location to a goal location on a surface. The path following process is based on concepts of places, path segments as connections between places, and the

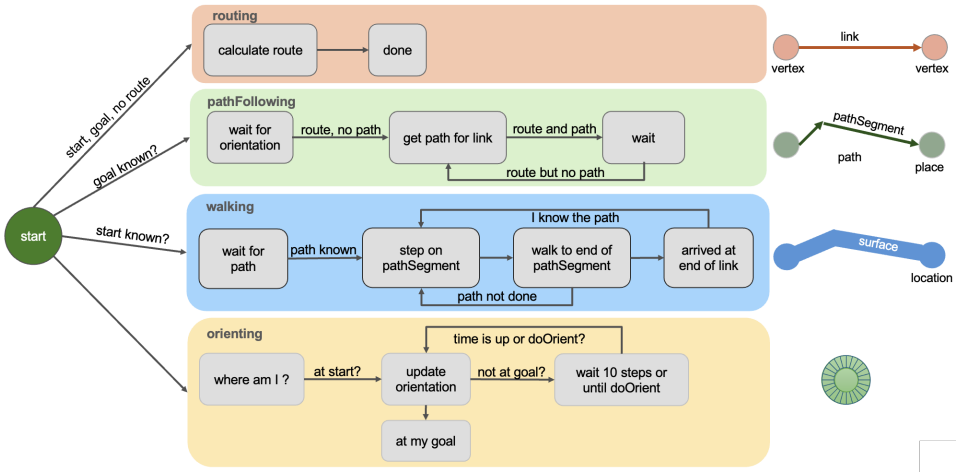


Figure 1 Conceptual model of behavioural processes and corresponding spatial representations.

notion of path as a connected series of path segments; finally, the routing process uses vertices, links, and the notion of a route. The path following process accesses metric information in its places and segments while the routing happens on a graph (seen as a purely topological structure). Locations are abstracted to places, some of which are abstracted to vertices. The surface in the walking process is simplified to a segment and a list of segments is abstracted to a link.

The conceptual model mostly follows the ideas of Downs and Stea [4] with the exception of the process of *route monitoring*. In this instance of our model, *route monitoring* consists of determining if the current position is consistent with being on the planned path. We suggest to integrate this process into the orienting process, with the monitoring being the comparison between the current and the planned orientation.

A key conceptual step is the stratification of behavioural processes. This separation is maintained by restricting interactions to shared agent or process states. Navigation is thus decomposed into four concurrent, theory-grounded sub-processes, forming the basis of our simulation.

Conceptualizing navigation as a set of concurrent, interdependent processes requires appropriate computational support. We implement the model in SeSAm[1], which natively supports concurrency through multiple statecharts, termed *reasoning engines*. Each reasoning engine encodes a single behavioural process at the agent level. Multiple engines can be active concurrently, enabling loosely coupled processes without enforced synchronization. The architecture is explicitly designed for agents engaging in concurrent activities (e.g., negotiating while executing a task).

The execution of concurrent processes is technically interleaved across atomic update cycles in randomized order, leaving process scheduling uncontrolled. This necessitates explicit coordination via shared informational states, aligning well with our design principle of process separation. The model code is available online on GitHub[2].

Figure 2 shows the four reasoning engines and the six shared state variables mediating their interactions. Processes influence one another indirectly via these variables (e.g., path-related information), avoiding centralized control. This preserves behavioural modularity.

[1] `http://simsesam.org/`
[2] `https://github.com/GeosavvyAgents/concurrentNavigationProcesses`

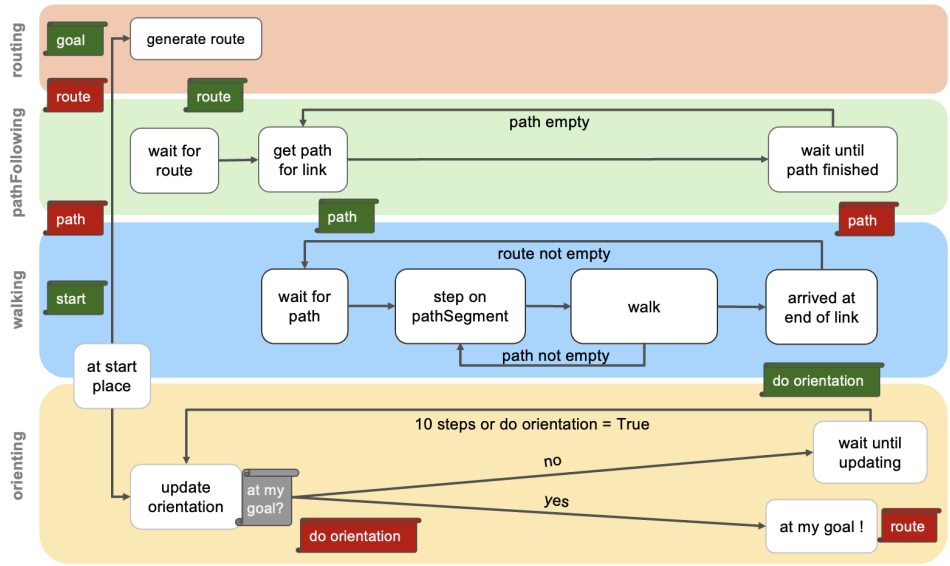


Figure 2 Four processes and their interaction with six state variables shown in green (state is true) and in orange (state is false).

Processes are conditionally activated and, once active, proceed concurrently. Each reasoning process can be replaced by alternative models or algorithms, provided they adhere to the same interface via shared state variables.

5 Discussion

The proposed framework reconceptualizes navigation as a set of concurrent processes coordinated through shared informational states, rather than sequential decision stages. This enables a more faithful representation of navigation as a continuous, situated activity in which perception, orientation, and movement interact dynamically. By decoupling processes and making dependencies explicit via shared state variables, the approach increases transparency compared to tightly integrated architectures.

The use of loosely coupled reasoning engines introduces a form of modularity that supports substitutability. Individual processes – such as routing or orientation – can be replaced by alternative models without requiring changes to the overall architecture, provided they conform to the shared state interface. This suggests a pathway toward reusable modelling patterns for navigation behaviour.

At the same time, this design introduces new challenges. The coordination of processes is no longer enforced by a central control structure but emerges from the correct design of shared state variables and enabling conditions. As a result, model validity depends critically on identifying appropriate informational states and their semantics, which remains a domain-specific and non-trivial task.

Finally, while the framework captures concurrency at the architectural level, the current implementation remains a proof of concept. The behavioural richness of individual processes is still limited, and their empirical grounding requires further investigation. Thus, the main contribution lies in the structuring of interaction rather than in the sophistication of the underlying behavioural models.

6 Conclusions and future work

This work-in-progress paper introduced a conceptual and computational framework for modelling human navigation as a set of interacting concurrent processes. By treating concurrency as a first-class modelling principle and coordinating processes through shared informational states, the approach provides an alternative to sequential and hierarchically structured agent architectures. The implementation in SeSAm demonstrates the feasibility of this design and highlights its potential for representing navigation as a dynamic and distributed activity.

Future work will proceed along three main directions. First, the individual reasoning processes will be extended with empirically grounded models from spatial cognition, enabling more realistic behaviour. Second, we aim to systematically identify and formalize the shared informational states required for coordinating concurrent processes, with the goal of developing reusable modelling patterns. Third, the framework will be evaluated through simulation experiments, including comparisons with sequential architectures to assess its impact on behavioural realism and explanatory power.

In the longer term, the proposed approach may contribute to a broader rethinking of agent-based models of spatial behaviour, in which decision-making is no longer conceptualized as a sequence of choices but as the emergent outcome of interacting processes operating across multiple temporal and representational scales.

References

1 Negar Alinaghi, Ioannis Giannopoulos, Markus Kattenbeck, and Martin Raubal. Decoding wayfinding: analyzing wayfinding processes in the outdoor environment. *Int. Journal of Geographical Information Science*, pages 1–31, March 2025. doi:10.1080/13658816.2025.2473599.

2 Eric Chown, Stephen Kaplan, and David Kortenkamp. Prototypes, location and associative networks (plan): Towards a unified theory of cognitive mapping. *Cognitive Science*, 19(1):1–52, 1995. doi:10.1016/0364-0213(95)90017-9.

3 Andrew T. Crooks, Christopher J. E. Castle, and Michael Batty. Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems*, 32(6):417–430, 2008. doi:10.1016/j.compenvurbsys.2008.09.004.

4 R.M. Downs and D. Stea. *Maps in Minds*. Harper & Row, New York, US, 1977.

5 Anna Charisse Farr, Tristan Kleinschmidt, Prasad Yarlagadda, and Kerrie Mengersen. Wayfinding: A simple concept, a complex process. *Transport Reviews*, 32(6):715–743, November 2012. doi:10.1080/01441647.2012.712555.

6 Gabriele Filomena and Judith A. Verstegen. Modelling the effect of landmarks on pedestrian dynamics in urban environments. *Computers, Environment and Urban Systems*, 86:101573, 2021. doi:10.1016/j.compenvurbsys.2020.101573.

7 M. Gath-Morad, E. Ronchi, and R. Lovreglio. A multi-agent approach for modelling pedestrian dynamics: A case study. In *Proceedings of the International Conference on Pedestrian and Evacuation Dynamics*, 2013.

8 Reginald G. Golledge. *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. Johns Hopkins University Press, Baltimore, 1999.

9 S. Gopal, Roberta L. Klatzky, and Thomas R. Smith. Navigator: A psychologically based model of environmental learning through navigation. In *Proceedings of the Spatial Information Theory Conference*, 1989.

10 D. Harel. Statecharts: a visual formalism for complex systems. *Science of Computer Programming*, 8(3):231–274, 1987. doi:10.1016/0167-6423(87)90035-9.

11 Barbara Hayes-Roth. A blackboard architecture for control. *Artificial Intelligence*, 26(3):251–321, July 1985. doi:10.1016/0004-3702(85)90063-3.

12 Dirk Helbing and Peter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995. doi:10.1103/PhysRevE.51.4282.

13 Peter M. Kielar and André Borrmann. Spice: a cognitive agent framework for computational crowd simulations in complex environments. *Autonomous Agents and Multi-Agent Systems*, 32(3):387–416, 2018. doi:10.1007/s10458-018-9383-2.

14 Angelika Kneidl, André Borrmann, and Dirk Hartmann. Generation and use of sparse navigation graphs for microscopic pedestrian simulation models. In *Advances in Intelligent Systems and Computing*, pages 189–198, 2012. doi:10.1007/978-3-642-29709-0_18.

15 L. Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison Wesley, Boston, 2002.

16 Thomas Madl, Ke Chen, Daniela Montaldi, and Robert Trappl. Computational cognitive models of spatial memory in navigation space: A review. *Neural Networks*, 65:18–43, 2015. doi:10.1016/j.neunet.2015.01.008.

17 D.R. Montello. Navigation. In P. Shah and A. Miyake, editors, *Handbook of visuospatial cognition*, pages 257–294. Cambridge University Press, Cambridge, 2005.

18 Mehdi Moussaïd, Mubbasir Kapadia, Trevor Thrash, Robert W. Sumner, Markus Gross, Dirk Helbing, and Christoph Hölscher. Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface*, 13(122):20160414, 2016. doi:10.1098/rsif.2016.0414.

19 Christoph Oechslein, Franziska Klügl, Rainer Herrler, and Frank Puppe. Uml for behavior-oriented multi-agent simulations. In Barbara Dunin-Keplicz and Edward Nawarecki, editors, *From Theory to Practice in Multi-Agent Systems*, pages 217–226, Berlin, Heidelberg, 2002. Springer. doi:10.1007/3-540-45941-3_23.

20 A. Omicini, F. Zambonelli, and M. Klusch, editors. *Coordination of Internet Agents: Models, Technologies, and Applications*. Springer, 2014.

21 A. S. Rao and M. P. Georgeff. BDI agents: From theory to practice. In *Proc. of the 1st Int. Conf. on Multiagent Systems, ICMAS95*, pages 312–319, 1995.

22 W. Reisig. *Understanding Petri Nets: Modeling Techniques, Analysis Methods, Case Studies*. Springer, Berlin, Heidelberg, 2013.

23 S. Russell and P. Norvig. *Artificial Intelligence – A Modern Approach*. Pearson, Harlow, 4 edition, 2022.

24 Angela Schwering, Jakub Krukar, Rui Li, Vanessa Joy Anacta, and Stefan Fuest. Wayfinding Through Orientation. *Spatial Cognition & Computation*, 17(4):273–303, October 2017. doi:10.1080/13875868.2017.1322597.

25 Sebastian Thrun. Robotic mapping: A survey. In *Exploring Artificial Intelligence in the New Millennium*, pages 1–35. Morgan Kaufmann, 2002.

26 Sabine Timpf and Werner Kuhn. Granularity Transformations in Wayfinding. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl Friedrich Wender, editors, *Spatial Cognition III, Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning*, Lecture Notes in Computer Science, pages 77–88. Springer, 2003. doi:10.1007/3-540-45004-1_5.

27 Sabine Timpf, Gary S. Volta, David W. Pollock, and Max J. Egenhofer. A conceptual model of wayfinding using multiple levels of abstractions. In A.U. Frank, I. Campari, and U Formentini, editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 of *LNCS*, pages 348–367. Springer, Heidelberg Berlin, 1992. doi:10.1007/3-540-55966-3_21.

28 Paul M. Torrens. Moving agent pedestrians through space and time. *Annals of the Association of American Geographers*, 102(1):35–66, 2012. doi:10.1080/00045608.2011.595658.

29 M. van Steen and A. Tanenbaum. *Distributed Systems*. distributed.systems.net, 4th edition, 2023.