

Enriching Location Representation with Detailed Semantic Information

Junyuan Liu ✉ 

SpaceTimeLab, University College London, UK

Xinglei Wang ✉ 

SpaceTimeLab, University College London, UK

Tao Cheng¹ ✉ 

SpaceTimeLab, University College London, UK

Abstract

Spatial representations that capture both structural and semantic characteristics of urban environments are essential for urban modeling. Traditional spatial embeddings often prioritize spatial proximity while underutilizing fine-grained contextual information from places. To address this limitation, we introduce **CaLLiPer+**, an extension of the CaLLiPer model that systematically integrates Point-of-Interest (POI) names alongside categorical labels within a multimodal contrastive learning framework. We evaluate its effectiveness on two downstream tasks – land use classification and socioeconomic status distribution mapping – demonstrating consistent performance gains of 4% to 11% over baseline methods. Additionally, we show that incorporating POI names enhances location retrieval, enabling models to capture complex urban concepts with greater precision. Ablation studies further reveal the complementary role of POI names and the advantages of leveraging pretrained text encoders for spatial representations. Overall, our findings highlight the potential of integrating fine-grained semantic attributes and multimodal learning techniques to advance the development of urban foundation models.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Knowledge representation and reasoning

Keywords and phrases Location Embedding, Contrastive Learning, Pretrained Model

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.3

1 Introduction

Spatial representations form the backbone of urban analysis, serving as essential tools for understanding and modeling complex urban systems. They underpin various applications, including urban functional distribution mapping [9, 10], land use classification [12], socioeconomic indicator estimation [11], future visitor prediction [5], and next-location prediction [8]. Traditional approaches typically encode locations as numeric coordinates or rely on spatial proximity [14, 15, 30], effectively capturing physical distance and structure. However, they often fail to capture the intricate functional interdependencies between places that drive urban dynamics.

In contrast, “patial” concepts emphasize the additional layers of meaning that humans ascribe to spaces, interpreting them through social, cultural, and functional attributes [7]. Point-of-Interest (POI) data offers a practical entry point for these attributes, as it couples spatial coordinates with descriptive names and labels. Such semantic information elucidates how different places function and interact within the broader urban landscape. Nevertheless, many existing embedding methods continue to emphasize spatial distance or simple categorical labels [9, 10, 28, 30, 31], underutilizing POI data’s finer-grained insights.

¹ Corresponding author



© Junyuan Liu, Xinglei Wang, and Tao Cheng;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O’Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 3; pp. 3:1–3:15



Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Recent innovations in deep learning and natural language processing [17, 19, 4] facilitate richer semantic alignments within spatial data. Notably, multimodal contrastive learning [22] has proven effective in aligning geographic coordinates with textual descriptions, thereby enhancing the semantic depth of spatial embeddings. A prime example is CaLLiPer [27], which aligns POI types with spatial coordinates to yield improvements in downstream tasks. However, CaLLiPer treats POI types as broad categorical labels, potentially overlooking the granular detail contained in POI names. Such names often provide specific and context-rich information, ranging from “Starbucks Coffee” to “John’s Hardware Store,” which can further enrich location understanding and distinguish unique POIs. Yet, the systematic integration of POI names into general-purpose spatial embeddings through multimodal contrastive learning remains underexplored. Addressing this gap is crucial for fully capturing the nuanced semantics of urban environments and advancing more comprehensive urban representation models.

To enhance spatial embeddings with richer semantic detail, we incorporate POI names alongside type labels into a multimodal contrastive learning framework. Building on the original CaLLiPer model, we propose an extended version called CaLLiPer+. We evaluate its effectiveness in two downstream tasks – Land Use Classification (LUC) and Socioeconomic Status Distribution Mapping (SDM) – as well as in an additional location retrieval task.

Our contributions are as follows:

1. We extend the CaLLiPer framework by incorporating POI names alongside type information, resulting in a unified model, CaLLiPer+ (§3).
2. We evaluate the enriched semantic representation on two downstream tasks, showing consistent performance gains of 4% to 11% over POI-type-only models (§5.1).
3. We conduct retrieval experiments to assess the model’s ability to capture urban concepts, and show that enriched semantics and advanced text encoders lead to better conceptual understanding (§5.2).
4. We demonstrate the effectiveness of contrastive learning with a pretrained encoder for location representation, and highlight the potential of the resulting embeddings for downstream applications (§6).

2 Related Work

2.1 Word Embeddings and Sentence Embeddings

The advancement of natural language processing (NLP) has led to powerful embedding techniques that transform textual data into high-dimensional vector spaces, enabling machines to better process and understand linguistic semantics. Early word embedding models such as Word2Vec [16] and GloVe [21] revolutionized NLP by capturing semantic relationships between words based on their co-occurrence in large text corpora.

Building on these foundational methods, sentence embedding models like Sentence-BERT [17] and SimCSE [6] were developed to generate dense representations of entire phrases or sentences while preserving contextual nuances. More recently, large language models (LLMs) such as BERT [3], GPT [23, 2, 20], and LLaMA [25] have further enhanced text embedding capabilities, facilitating sophisticated semantic extraction across various textual contexts, including POI descriptions and names.

These advancements in NLP offer new opportunities to incorporate linguistic semantics into geospatial models, enabling the embedding of POI names and descriptions to enrich spatial representations beyond purely numerical features.

2.2 Spatial Embeddings with POIs

Spatial embedding techniques aim to encode geographic entities into vector spaces, capturing their spatial and functional relationships. POI data, which contains both geographic coordinates and semantic attributes, has been widely utilized in urban studies for tasks such as land use classification, urban function recognition, and socioeconomic mapping.

Early approaches to spatial embeddings primarily leveraged POI categories to model urban entity co-occurrence. Yao et al. [30] proposed a method that traversed POIs within a geographic region using shortest-path algorithms to extract co-occurrence patterns. Place2Vec [28] applied a K-nearest neighbor (KNN) sampling strategy with distance decay to model spatial proximity, while Doc2Vec [18] treated urban regions as documents and POIs as words, learning region embeddings based on the co-occurrence of POI categories within predefined spatial boundaries. These methods effectively captured the functional composition of urban spaces but treated POIs as categorical variables, overlooking their individual characteristics and richer semantic meanings.

To provide more distinguishing information for individual POIs, recent methods have explored integrating additional semantic attributes into spatial embeddings. Huang et al. [9] introduced the Semantics-Preserved POI Embedding (SPPE) model, which incorporates both spatial co-occurrence patterns and categorical semantics to enhance the representation of POI distributions. Similarly, HGI [10] employed hierarchical graph-based embeddings to capture multi-level semantic relationships among POIs, urban regions, and cities. While these methods improved the semantic richness of spatial representations, they still primarily rely on categorical classifications and predefined spatial structures, limiting their adaptability to diverse urban environments.

Existing methods for spatial embeddings primarily aggregate POI information within predefined regions or construct complex spatial contexts to infer urban functions. These approaches often rely on indirect or coarse-grained representations. With the growing availability of detailed POI datasets and advances in NLP, a more direct and efficient approach is to embed individual POIs by leveraging their inherent semantic information, such as names, which provide fine-grained functional and cultural context.

2.3 Multimodal Contrastive Learning for Geospatial Data

Multimodal contrastive learning has recently gained traction as an effective method for aligning heterogeneous data sources, enabling the integration of spatial coordinates with diverse information. This approach leverages contrastive objectives to maximize similarity between aligned data pairs (e.g., a location and its textual description) while distinguishing them from unrelated samples.

UrbanCLIP [29] proposed a pre-training approach for urban region representation by generating textual descriptions for satellite images using large language models and training an image encoder via a CLIP-like framework. Similarly, GeoCLIP [26] and SatCLIP [13] extended contrastive learning to geospatial data by aligning satellite imagery with geographic coordinates, supporting tasks such as geo-localization and environmental monitoring. The CaLLiPer model [27] advanced this concept by aligning POI type semantics with spatial coordinates through multimodal contrastive learning, demonstrating improved performance in land use classification and socioeconomic status mapping.

Despite these advances, existing models primarily focus on solely POI type or complex visual data, overlooking the potential benefits of simply incorporating distinguishing semantics of POI names into contrastive learning settings, which contain rich, context-specific

information that can enhance the semantic depth of spatial embeddings, offering more nuanced insights into urban functions and structures. The underutilization of POI names in multimodal frameworks is still a significant gap in current geospatial representation learning research.

3 Methodology

3.1 Overview

This study builds upon the CaLLiPer framework [27], a multimodal contrastive learning model designed to align spatial coordinates with semantic information extracted from POI data. While the core architecture remains consistent with CaLLiPer, we introduce a key modification: the integration of POI names into the textual descriptions, enriching the semantic representation of urban spaces.

Figure 1 illustrates the overall architecture, which consists of three key components: a location encoder, a text encoder, and a projection layer. These components are jointly optimized using a contrastive learning objective to align spatial and semantic information effectively.

Location encoder. The location encoder maps spatial coordinates into a continuous vector space. It applies a positional encoding function to transform raw geographic coordinates into structured representations, followed by a fully connected neural network to generate location embeddings. In this work, we apply the Grid [14] positional encoding function.

Text encoder. The text encoder is a frozen pretrained embedding model, such as Sentence-BERT [17], LLaMA [25], or GPT [20], which generates semantic embeddings from the enriched POI descriptions. By incorporating POI names alongside categorical information, it captures more nuanced semantic details, improving the discriminative power of the embeddings.

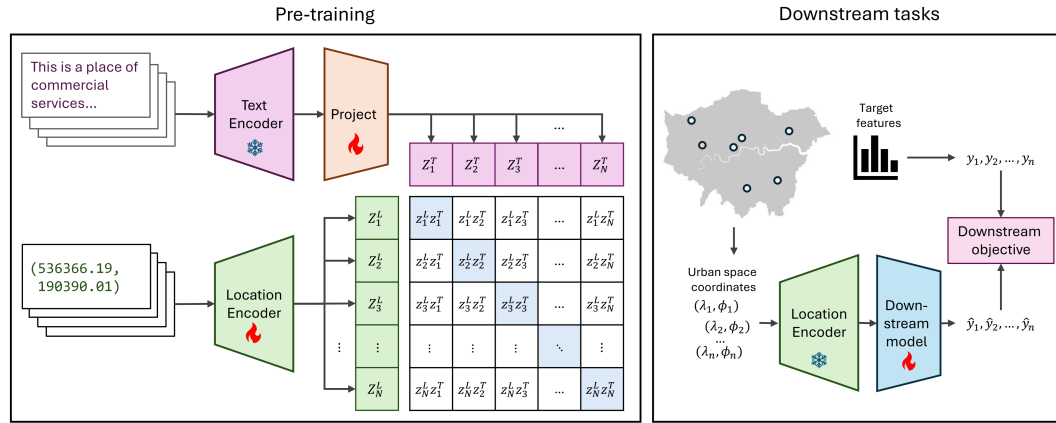
Projection layer. To facilitate direct comparison between spatial and textual embeddings, a linear projection layer maps both of them into a common vector space of dimension d . This projection ensures compatibility between modalities, enabling effective contrastive learning.

3.2 Enriching POI Descriptions with Names

In the original CaLLiPer model, POI semantics are represented solely by two levels of categorical labels from the Ordnance Survey. While effective for generalizing urban functions, this approach overlooks the rich, context-specific information embedded in POI names. Names often convey distinctive characteristics, such as cultural significance, brand identity, or specialized services, which are not captured by generic type labels. For instance, “McDonald’s” may evoke a different functional connotation compared to a generic “restaurant,” particularly in terms of cuisine style or consumption level.

To address this limitation, we extend the POI descriptions by integrating names directly into the semantic representation. For each POI p_i , we construct a combined description d_i that incorporates the name n_i , the first-level category t_{1i} , and the second-level class t_{2i} using a templated format designed to enhance the model’s understanding of the spatial context:

$$d_i = \text{Template}(n_i, t_{1i}, t_{2i}) = \text{“A place of } [t_{2i}], \text{ a type of } [t_{1i}], \text{ named } [n_i].\text{”} \quad (1)$$



■ **Figure 1** Architecture of the CaLLiPer+ model [27]. POI names are incorporated into the textual descriptions processed by the text encoder, enhancing the semantic richness of the spatial embeddings.

This enriched template ensures that the text encoder can capture both general category information and the specific nuances associated with individual POIs. By incorporating POI names, the model captures finer-grained semantic details that improve its ability to differentiate between places within the same category. This includes recognizing brand prestige (e.g., “Hilton Hotel” vs. “Budget Inn”), specific function within the same type (e.g., “The British Museum” vs. “National Gallery”), and scale or exclusivity (e.g., “local farm market” vs. “Harrods”). This richer semantic embedding enhances the model’s capacity to represent the diversity and complexity of urban environments more accurately.

3.3 Multimodal Contrastive Learning Framework

The multimodal contrastive learning framework aligns spatial coordinates with detailed textual semantics in a shared embedding space. The goal is to ensure that a POI’s spatial representation is closely aligned with its semantic description, while unrelated pairs are pushed apart.

Each POI is represented by two embeddings:

$$z_i^{(s)} = f_s(\mathbf{x}_i) \quad (\text{spatial embedding}) \quad (2)$$

$$z_i^{(p)} = W_t f_t(d_i) \quad (\text{textual embedding with name and type}) \quad (3)$$

where f_s is the spatial encoder that transforms the geographic coordinates $\mathbf{x}_i$ into a vector representation, and f_t is a pretrained text encoder that processes the enriched POI descriptions d_i , followed by a projection layer W_t to align the dimension with spatial embedding. The inclusion of POI names in d_i ensures that the text embeddings capture both high-level categorical information and fine-grained, context-specific details.

Contrastive learning objective. The alignment between spatial and textual embeddings is achieved using the InfoNCE loss [22], which encourages positive pairs (i.e., a POI’s location and its enriched description) to be similar, while pushing apart negative pairs (i.e., mismatched locations and descriptions). The loss is defined as:

$$\mathcal{L} = -\frac{1}{2N} \left[\sum_{i=1}^N \log \frac{\exp(z_i^{(s)} \cdot z_i^{(p)} / \tau)}{\sum_{j=1}^N \exp(z_i^{(s)} \cdot z_j^{(p)} / \tau)} + \sum_{i=1}^N \log \frac{\exp(z_i^{(p)} \cdot z_i^{(s)} / \tau)}{\sum_{j=1}^N \exp(z_i^{(p)} \cdot z_j^{(s)} / \tau)} \right], \quad (4)$$

where $\cdot$ denotes cosine similarity between embeddings, and τ is a temperature parameter that controls the sharpness of the distribution. This symmetric loss is applied to both spatial-to-textual and textual-to-spatial alignment, ensuring consistent alignment of embeddings from both modalities.

Advantages of enriched semantics. Incorporating POI names into the contrastive framework enhances the model’s ability to capture fine-grained urban semantics. The enriched descriptions provide the following benefits:

- Improved discrimination: The model can better differentiate between places within the same category by leveraging unique names.
- Context awareness: Names often imply cultural, historical, or functional context, enriching the model’s understanding of urban environments.
- Enhanced transferability: The enriched embeddings generalize better across diverse tasks.

In summary, our approach enhances the original CaLLiPer framework by incorporating POI names into the textual descriptions, leading to richer, more discriminative spatial embeddings through multimodal contrastive learning.

4 Experiments

4.1 Experimental Setup

To evaluate the impact of incorporating POI names into the spatial-semantic embeddings, we conducted experiments on two urban analytics tasks: Land Use Classification (LUC) and Socioeconomic Status Distribution Mapping (SDM). Additionally, we performed location retrieval to observe the model’s ability to capture high-level urban concepts.

4.2 Datasets

Point-of-Interest data. We use POI data from the Ordnance Survey via Digimap ², covering the Greater London area. The dataset contains approximately 340,000 POIs, each with geographic coordinates, a name, and categorical labels. POIs are classified into a hierarchical taxonomy. These data provide detailed spatial and semantic insights into London’s urban environment.

Land use data. We obtain land use data from the Verisk National Land Use Database ³, which provides high-resolution classification of land use types. The dataset includes ten primary land use categories. To create the evaluation dataset, we sample locations with a 200-meter radius buffer, ensuring balanced representation across categories.

Socioeconomic data. We obtain socioeconomic data from the Office for National Statistics (ONS) 2021 Census ⁴, specifically the National Statistics Socioeconomic Classification (NS-SeC). This dataset provides a detailed classification of socioeconomic status based on employment type, occupational hierarchy, and educational attainment. The data are aggregated at the Lower-layer Super Output Area (LSOA) level, encompassing 4,994 LSOAs across London. Each LSOA contains proportions of 1000 to 3000 residents within different occupational classes.

² <https://digimap.edina.ac.uk/>

³ <https://digimap.edina.ac.uk/roam/map/verisk>

⁴ <https://www.ons.gov.uk/>

4.3 Baselines

To assess the effectiveness of our enhanced model, CaLLiPer+, we compare it against the following baselines:

- **TF-IDF** [24]: A term frequency-inverse document frequency model that represents each region based on the POI categories within it.
- **LDA** [1]: A probabilistic topic modeling approach that infers latent topics from POI distributions, capturing urban functional structures through topic-word distributions.
- **Place2Vec** [28]: A spatial embedding model that learns representations of POIs based on their spatial co-occurrence, modeling functional similarity through a skip-gram framework.
- **Doc2Vec** [18]: A document embedding approach that treats urban regions as documents composed of POI categories, learning region representations through unsupervised learning.
- **SPPE** [9]: A semantics-preserving POI embedding method that captures spatial co-occurrence patterns and topological structures of POIs through a graph-based approach.
- **Space2Vec** [14]: A geospatial representation learning model that encodes locations through positional encoding and neural networks, learning embeddings directly from spatial coordinates.
- **CaLLiPer** [27]: The original multimodal contrastive learning model, which encodes POI categories as textual descriptions but does not incorporate POI names.

4.4 Downstream Tasks and Evaluation Metrics

We evaluate the learned spatial representations on LUC and SDM tasks. To systematically analyze the effectiveness of the learned embeddings, we employ two types of downstream models: (1) a linear model, implemented as a single-layer neural network, testing the raw expressiveness of the embeddings, and (2) a nonlinear model, implemented as a multi-layer perceptron (MLP) with a single hidden layer to capture more complex relationships.

Land use classification is a multi-class classification task that predicts the land use type of a given spatial unit based on its learned representation. We train classifiers using both a linear model and a nonlinear model and evaluate performance using:

- Precision, recall, and F1 score: These metrics are macro-averaged across classes, providing a balanced evaluation of classification performance. Higher values indicate better performance.

Socioeconomic status distribution mapping is a regression task that estimates the occupational composition of urban regions using the learned embeddings. The model predicts the proportion of residents in different socioeconomic categories at the LSOA level. We train both a linear model and a nonlinear model to compare their effectiveness. Performance is evaluated using:

- L1 distance: Measures the absolute difference between predicted and actual socioeconomic distributions.
- Chebyshev distance: Captures the maximum absolute deviation between predicted and actual distributions.
- Kullback-Leibler (KL) divergence: Evaluates the difference between the predicted and actual probability distributions, indicating how well the model captures the socioeconomic structure.

By testing the embeddings across both classification and regression tasks, and using both linear and nonlinear models, we assess their generalizability and effectiveness in capturing the information of urban environments.

4.5 Implementation Details

All models were implemented using PyTorch and trained on a machine equipped with an NVIDIA A6000 GPU. The text encoder was based on Sentence-BERT by default, which processed the enriched POI descriptions. The spatial encoder followed the same architecture as in CaLLiPer [27], using a fully connected residual network with 128-dimensional embeddings. The training process adopted a grid search approach to tune hyperparameters, resulting in a batch size of 128, a learning rate of 0.0001, and a temperature parameter of 0.07. The optimizer was Adam. The models were trained for 100 epochs with early stopping based on validation loss, and each downstream task experiment was repeated five times with different random seeds to ensure robustness. The reported results represent the mean performance across these runs.

4.6 Location Retrieval

We observe the model’s ability to retrieve urban concepts based on semantic queries. This task shows how well the learned embeddings capture urban concepts by matching textual embeddings to spatial embeddings.

Given a natural language query, we compute its embedding using a pretrained language model. We use two text encoding approaches: (1) a Sentence-Transformers model (all-MiniLM-L6-v2), which generates sentence embeddings via mean pooling over contextualized token embeddings, and (2) an OpenAI GPT-based embedding model (text-embedding-3-small), which produces a high-dimensional representation of the query and is subsequently projected into a 128-dimensional space for compatibility with the learned spatial embeddings.

The model then retrieves the most relevant locations by computing cosine similarity between the query embedding and the location embeddings of urban regions. To assess retrieval effectiveness, we visualize the top-ranked locations using geospatial maps, highlighting areas with the highest similarity to the input query.

4.7 Ablation Study

To evaluate the impact of different semantic components and text encoders, we conduct an ablation study with four model variants:

- **CaLLiPer+ GPT**: A variant that replaces the sentence transformer with GPT (text-embedding-3-small), examining the effect of a text embedding from LLM. For fairness, we only use the first 384 dimensions of the text embedding, which is the same as the default sentence transformer.
- **CaLLiPer+**: The default enhanced model that integrates both POI names and types, using a sentence transformer (all-MiniLM-L6-v2).
- **CaLLiPer+ w/o type**: A variant that removes POI types, using only POI names for textual representation.
- **CaLLiPer**: A variant that excludes POI names and relies only on POI types, which is the original CaLLiPer.

We evaluate these models on the LUC and SDM tasks. The primary metrics used are F1 score for classification and KL divergence for regression-based analysis. The results are summarized in Figure 4.

■ **Table 1** Performance comparison on the LUC task. The best and second-best performances are marked in **bold** and underlined, respectively. For better readability, all metrics are scaled by a factor of 10^2 .

Model	Linear			MLP		
	Precision $\uparrow$	Recall $\uparrow$	F1 Score $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1 Score $\uparrow$
Random	9.6 ± 0.7	10.3 ± 0.5	9.7 ± 0.5	8.8 ± 1.3	10.3 ± 0.3	9.0 ± 0.3
TF-IDF	31.5 ± 0.4	32.2 ± 0.2	31.3 ± 0.3	31.8 ± 0.6	33.3 ± 0.5	31.7 ± 0.6
LDA	30.8 ± 0.3	29.1 ± 0.2	28.4 ± 0.2	31.5 ± 1.1	30.4 ± 0.7	29.2 ± 0.9
Place2Vec	30.9 ± 0.8	26.1 ± 0.7	26.3 ± 0.7	35.1 ± 1.2	32.7 ± 1.0	32.4 ± 1.2
Doc2Vec	32.4 ± 0.4	28.2 ± 0.1	28.0 ± 0.1	34.9 ± 0.9	33.8 ± 0.5	32.7 ± 0.6
SPPE	30.5 ± 0.4	27.0 ± 0.2	26.6 ± 0.2	34.5 ± 0.9	32.9 ± 0.7	32.2 ± 0.5
HGI	33.0 ± 0.5	30.0 ± 0.6	29.9 ± 0.6	33.6 ± 0.5	32.0 ± 0.9	31.6 ± 0.7
Space2Vec	28.6 ± 0.6	28.5 ± 0.8	27.4 ± 0.7	29.6 ± 0.6	28.9 ± 0.5	27.8 ± 0.3
CaLLiPer	36.5 ± 0.6	35.3 ± 0.2	34.6 ± 0.3	37.7 ± 0.8	35.5 ± 0.8	34.6 ± 0.8
CaLLiPer+	<u>37.5 ± 0.7</u>	<u>35.5 ± 0.5</u>	<u>35.2 ± 0.6</u>	<u>40.0 ± 0.4</u>	<u>36.0 ± 0.5</u>	<u>36.6 ± 0.5</u>
CaLLiPer+GPT	40.5 ± 0.6	36.7 ± 0.2	36.8 ± 0.3	41.3 ± 0.7	37.8 ± 0.4	37.6 ± 0.3

■ **Table 2** Performance comparison on the SDM task. The best and second-best performances are marked in **bold** and underlined, respectively. For better readability, all metrics are scaled by a factor of 10^2 .

Model	Linear			MLP		
	L1 $\downarrow$	Chebyshev $\downarrow$	KL $\downarrow$	L1 $\downarrow$	Chebyshev $\downarrow$	KL $\downarrow$
Random	30.31 ± 0.03	9.25 ± 0.01	7.73 ± 0.01	31.40 ± 0.22	9.55 ± 0.11	8.21 ± 0.14
TF-IDF	24.79 ± 0.04	7.43 ± 0.01	5.36 ± 0.01	24.36 ± 0.15	7.28 ± 0.05	5.20 ± 0.04
LDA	26.14 ± 0.01	7.84 ± 0.00	5.87 ± 0.00	25.85 ± 0.14	7.77 ± 1.12	5.80 ± 0.72
Place2Vec	23.47 ± 0.09	6.94 ± 0.02	4.81 ± 0.02	22.81 ± 0.06	6.81 ± 0.01	4.61 ± 0.02
Doc2Vec	24.01 ± 0.07	7.15 ± 0.02	4.99 ± 0.02	23.10 ± 0.19	6.89 ± 0.06	4.75 ± 0.08
SPPE	24.32 ± 0.16	7.24 ± 0.06	5.11 ± 0.06	23.63 ± 0.19	7.04 ± 0.06	4.91 ± 0.07
HGI	23.28 ± 0.08	6.93 ± 0.02	4.79 ± 0.03	22.73 ± 0.05	6.80 ± 0.02	4.60 ± 0.02
Space2Vec	25.13 ± 0.15	7.56 ± 0.04	5.65 ± 0.06	23.55 ± 0.20	7.12 ± 0.09	5.00 ± 0.08
CaLLiPer	21.63 ± 0.04	6.55 ± 0.05	4.26 ± 0.01	20.52 ± 0.14	6.24 ± 0.03	3.90 ± 0.06
CaLLiPer+	<u>20.87 ± 0.02</u>	<u>6.35 ± 0.01</u>	<u>3.98 ± 0.01</u>	<u>19.85 ± 0.19</u>	<u>6.02 ± 0.06</u>	<u>3.63 ± 0.07</u>
CaLLiPer+GPT	20.26 ± 0.03	6.09 ± 0.01	3.74 ± 0.01	19.38 ± 0.02	5.83 ± 0.04	3.47 ± 0.01

5 Results and Analysis

5.1 Performance on Downstream Tasks

Tables 1 and 2 summarize the results for LUC and SDM tasks. Across both tasks, multimodal contrastive learning models outperform traditional methods, demonstrating the effectiveness of integrating spatial and semantic information. Baseline models such as TF-IDF and LDA rely on aggregated POI type distributions within regions, limiting their ability to capture fine-grained relationships between locations. While methods like Place2Vec and Doc2Vec improve upon this by incorporating spatial co-occurrence structures, their reliance on unsupervised embedding techniques without explicit spatial-semantic alignment leads to weaker performance. In contrast, CaLLiPer and its extensions, which align POI-based textual representations with spatial coordinates, consistently achieve better results, confirming the advantages of multimodal contrastive learning.

Additionally, CaLLiPer+ achieves superior and more stable performance across all metrics. In LUC, CaLLiPer+ consistently outperforms the original CaLLiPer model, achieving higher precision, recall, and F1 scores across both linear and MLP classifiers. This demonstrates that integrating POI names alongside type-based descriptions enriches the model's semantic understanding of urban space, allowing for better land use classification. A similar trend is observed in SDM, where CaLLiPer+ further reduces errors across all three evaluation metrics, suggesting that POI names provide valuable contextual information for modeling socioeconomic distributions. Notably, CaLLiPer+ GPT achieves the best performance across both tasks, reinforcing the importance of using more powerful text encoders for spatial representation learning.

Third, the improvements observed with MLP over the linear model suggest that the learned embeddings still contain complex, non-linear relationships that can be further leveraged by downstream tasks. While baseline models such as TF-IDF and LDA show limited gains with MLP, indicating that their representations are mostly exhausted by simple classifiers, CaLLiPer-based models still exhibit a more notable performance boost. CaLLiPer+ effectively aligns spatial and semantic information, and the embeddings still retain structured patterns that require more expressive models to fully exploit, highlighting the depth and richness of the learned representations.

These findings highlight the advantages of incorporating both POI names and stronger text embedding models for geospatial representation learning, improving the model's ability to capture complex urban semantics across diverse tasks.

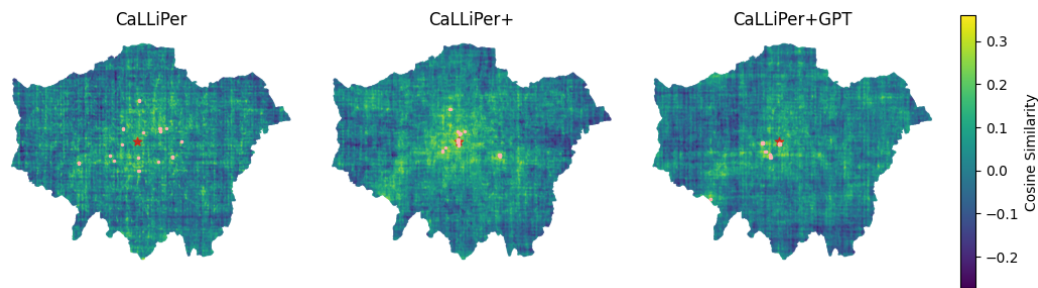
5.2 Location Retrieval

Location retrieval evaluates the model's ability to associate spatial embeddings with meaningful semantic queries, including specific place names and abstract urban concepts. The results, shown in Figures 2 and 3, illustrate how different model variants respond to retrieval tasks.

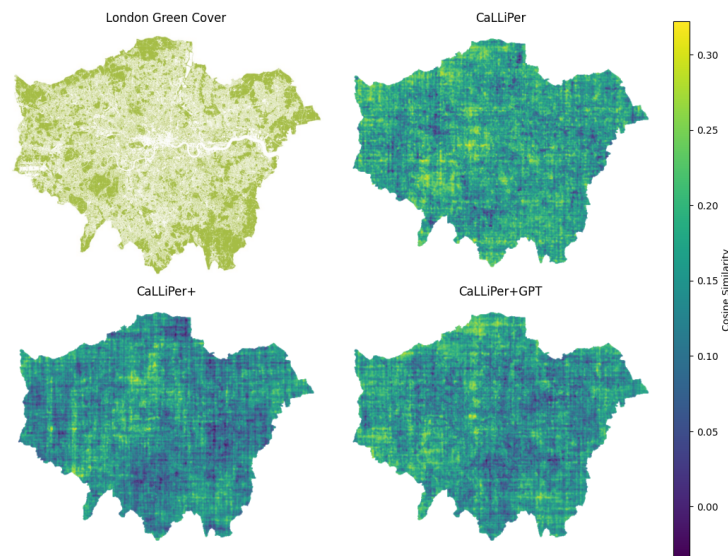
First, using POI names directly for retrieval demonstrates that including POI names in the text encoder significantly improves the model's ability to locate specific places. In Figure 2, models that incorporate POI names (CaLLiPer+ and CaLLiPer+GPT) produce more precise and concentrated retrieval results compared to the original CaLLiPer model, which relies solely on categorical types. The use of a more powerful text encoder, such as GPT embeddings in CaLLiPer+GPT, further enhances localization, leading to more accurate spatial responses.

Second, for high-level conceptual retrieval, such as identifying regions characterized by abstract urban concepts (e.g., green cover), the inclusion of POI names introduces both benefits and challenges. As seen in Figure 3, models that incorporate POI names sometimes exhibit increased dispersion in similarity scores when handling broad, high-level concepts. This suggests that when the model's semantic understanding is insufficient, in such cases, additional name-based details can introduce ambiguity. However, when equipped with a more advanced text encoder (e.g., CaLLiPer+GPT), the model can effectively utilize this additional semantic information to establish clearer distinctions between different urban functions, demonstrating improved conceptual retrieval. This improvement can be attributed to GPT's ability to capture hierarchical urban concepts and their interconnections, enabling a more nuanced understanding of spatial semantics.

Overall, our results highlight the benefits of integrating POI names in location retrieval. Name-enhanced models improve direct place retrieval and, with sufficiently strong text encoders, also facilitate better discrimination of abstract spatial concepts.



■ **Figure 2** Similarity map for “The National Gallery.” The red star is the actual location of the target, and the yellow points are the top 30 similar locations.



■ **Figure 3** Similarity map for “A place of park or green cover.” The ground truth is based on green cover data from London DataStore ⁵.

5.3 Ablation Study Results

Figure 4 presents the results of our ablation study. Both POI names and types contribute to improving downstream tasks, as seen from the superior performance of CaLLiPer+ compared to CaLLiPer and CaLLiPer+ w/o type. This suggests that combining both sources of semantic information leads to more informative spatial representations.

Interestingly, even when POI types are removed (CaLLiPer+ w/o type), the model still outperforms CaLLiPer, indicating that POI names carry richer and more discriminative semantic details than type labels alone. This highlights the potential of leveraging fine-grained textual information like POI names in spatial embedding models.

Moreover, using a stronger text encoder (CaLLiPer+ GPT) further improves results across both tasks. The enhanced semantic representation from a large language model allows for a better understanding of the text concepts in urban semantics, reinforcing the importance of high-quality embeddings in geospatial contrastive learning.

⁵ <https://apps.london.gov.uk/green-cover>

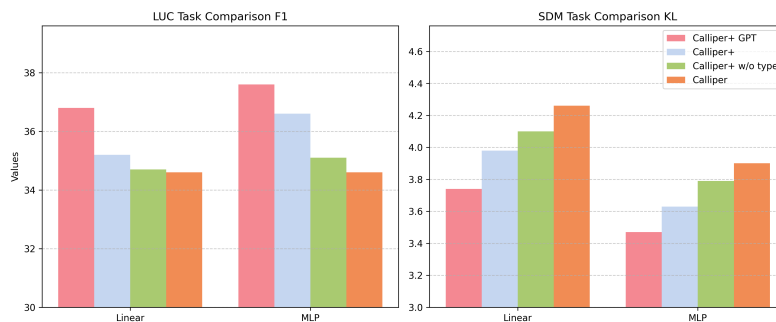


Figure 4 Ablation study results comparing model variations across LUC and SDM tasks. The left plot shows F1 score $\uparrow$ performance on LUC, while the right plot presents KL divergence $\downarrow$ results for SDM. All metrics are scaled by a factor of 10^2 .

6 Discussion and Conclusion

We explore the impact of integrating POI names into multimodal contrastive learning for spatial representation. By extending the CaLLiPer framework to incorporate both POI types and names, we introduce CaLLiPer+, which enhances the semantic richness of location embeddings. Our experiments across land use classification, socioeconomic status distribution mapping, and location retrieval reveal key insights into the role of enriched textual descriptions in geospatial learning.

Effectiveness of POI names in spatial representation. The combining of POI names with types in multi-modal contrastive learning improves downstream task performance consistently. POI names provide more specific and context-aware semantic signals, capturing fine-grained distinctions that categorical types alone may overlook. This effect is particularly evident in retrieval tasks, where name-enhanced models demonstrate greater precision in identifying specific locations.

Impact of text encoder strength. Using more advanced text embeddings, such as those from GPT-based models, further refines spatial representation. The CaLLiPer+ GPT model consistently outperforms others, suggesting that stronger language models contribute to a deeper understanding of urban semantics. This aligns with findings in location retrieval, where better text embeddings enable clearer conceptual differentiation, especially for high-level concepts.

Limitations and future work. The quality of spatial embeddings relies on the density and distribution of POIs across different urban areas. Regions with too sparse POI coverage may lead to less informative representations, limiting generalizability. Also, the information beyond the semantics still needs to be explored. Future work should incorporate additional modalities such as road networks, street-view imagery, and mobility patterns to enrich spatial information. Additionally, while our current downstream tasks provide initial validation, further research should explore a wider range of urban analytics applications and develop task-specific models that better leverage the structure of learned embeddings for improved adaptability and performance.

Conclusion. This work demonstrates that incorporating POI names into geospatial contrastive representation learning leads to improved performance in multiple urban analytics tasks. By aligning spatial and semantic information more effectively, CaLLiPer+ provides a more detailed and context-aware model for understanding urban environments. The effectiveness of semantic information highlights the potential of using pretrained multimodal models to generate enriched spatial embeddings in advancing urban intelligence.

References

- 1 David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. URL: <https://jmlr.org/papers/v3/blei03a.html>.
- 2 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 3 Jacob Devlin, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
- 4 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint*, 2024. [arXiv:2407.21783](https://arxiv.org/abs/2407.21783).
- 5 Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
- 6 Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021. doi:10.18653/V1/2021.EMNLP-MAIN.552.
- 7 Michael F. Goodchild. Platial. In *International Encyclopedia of Geography*, pages 1–5. Wiley, September 2020. doi:10.1002/9781118786352.wbieg2046.
- 8 Ye Hong, Yatao Zhang, Konrad Schindler, and Martin Raubal. Context-aware multi-head self-attentional neural network model for next location prediction. *Transportation Research Part C: Emerging Technologies*, 156:104315, 2023.
- 9 Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36(10):1905–1930, 2022. doi:10.1080/13658816.2022.2040510.
- 10 Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145, 2023.
- 11 Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- 12 Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3967–3974, 2019. doi:10.1609/AAAI.V33I01.33013967.
- 13 Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. doi:10.48550/arXiv.2311.17179.

- 14 Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020. URL: <https://openreview.net/forum?id=rJ1jdh4KDH>.
- 15 Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023.
- 16 Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint*, 2013. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- 17 Reimers Nils and Gurevych Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. In *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, 2019.
- 18 Haifeng Niu and Elisabete A Silva. Delineating urban functional use from points of interest data with neural network embedding: A case study in greater london. *Computers, Environment and Urban Systems*, 88:101651, 2021. doi:10.1016/J.COMPENVURBSYS.2021.101651.
- 19 OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022. Accessed: 2023-07-26.
- 20 OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi:10.48550/arXiv.2303.08774.
- 21 Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. doi:10.3115/V1/D14-1162.
- 22 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- 23 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 24 Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
- 25 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- 26 Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024.
- 27 Xinglei Wang, Tao Cheng, Stephen Law, Zichao Zeng, Lu Yin, and Junyuan Liu. Multi-modal contrastive learning of urban space representations from poi data. *Computers, Environment and Urban Systems*, 120:102299, 2025. doi:10.1016/J.COMPENVURBSYS.2025.102299.
- 28 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017. doi:10.1145/3139958.3140054.
- 29 Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017, 2024. doi:10.1145/3589334.3645378.

- 30 Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848, 2017. doi:10.1080/13658816.2016.1244608.
- 31 Wei Zhai, Xueyin Bai, Yu Shi, Yu Han, Zhong-Ren Peng, and Chaolin Gu. Beyond word2vec: An approach for urban functional region extraction and identification by combining place2vec and pois. *Computers, environment and urban systems*, 74:1–12, 2019. doi:10.1016/J.COMPENVURBSYS.2018.11.008.