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].

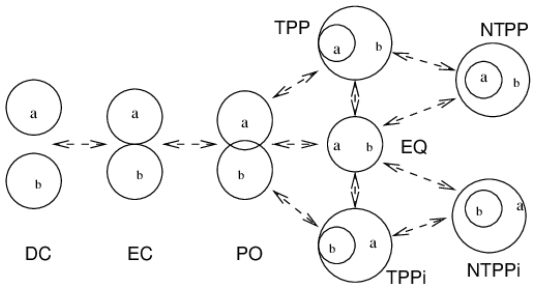

Figure 1 The eight relations of the RCC-8 calculus illustrated in 2D [11]: DC (Disconnected), EC (Externally Connected), PO (Partially Overlapping), TPP (Tangential Proper Part), NTPP (Nontangential Proper Part) and EQ (Equals); TPPi and NTPPi are the inverses of TPP and NTPP respectively since they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

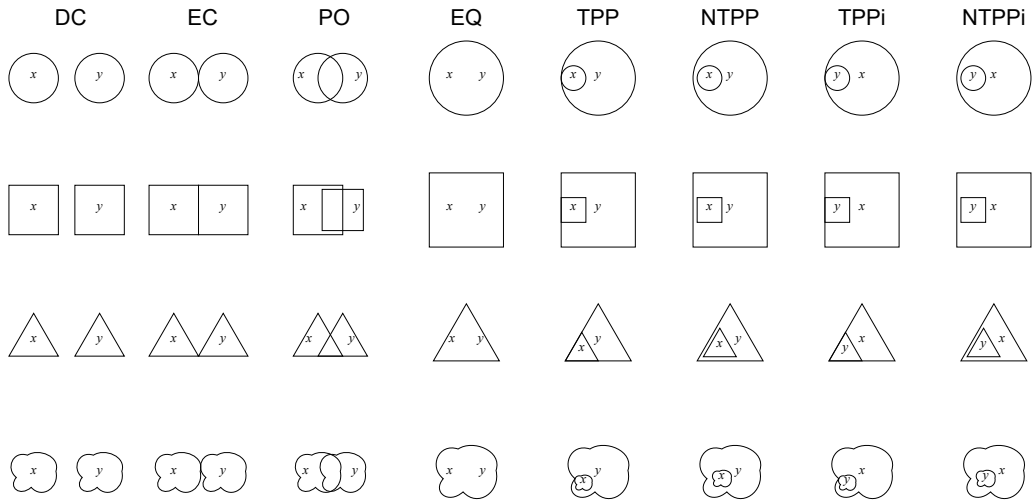

Figure 2 RCC-8 relation diagrams for circles, squares, triangles, and blobs. We use these diagrams in both SVG and PNG format to test the ability of VLMs to recognise the RCC-8 relations.

```
<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.

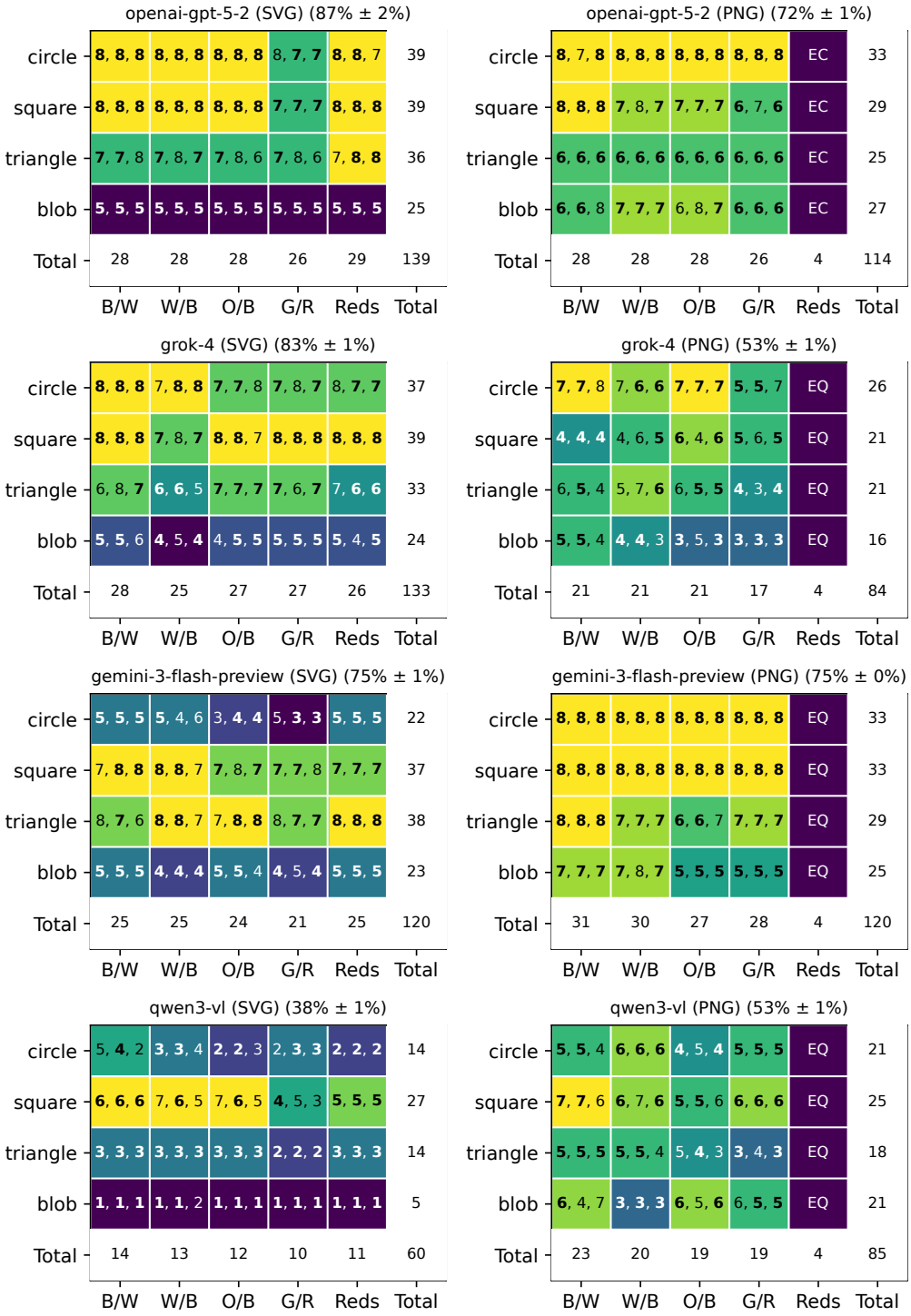

Figure 3 Vision-Language Model relation diagram recognition comparing shape and colour combinations for SVG and PNG by model. Each model run has 160 questions (eight relations × five colour combinations × four shapes), so the score in each cell is out of 8, with n=3 repeats. Rows and columns are ordered 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.

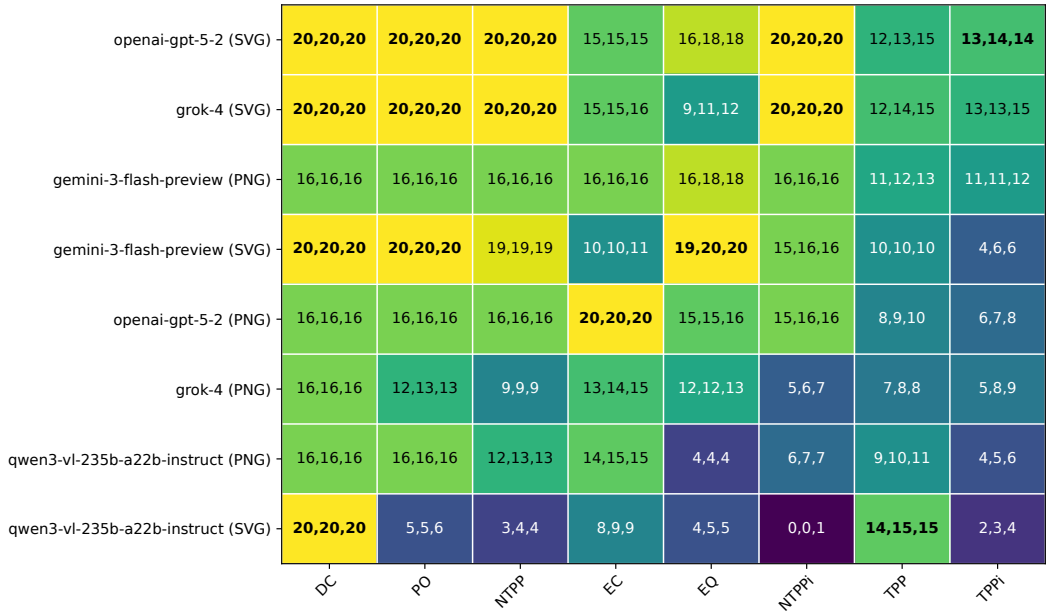

Figure 4 Vision-Language Model relation diagram recognition for each of the RCC-8 relations. Each model run has 160 questions (eight relations × five colour combinations × four shapes), and so each score is out of 20, with n=3 repeats. Rows and columns are ordered by the sum of the median scores, so 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*

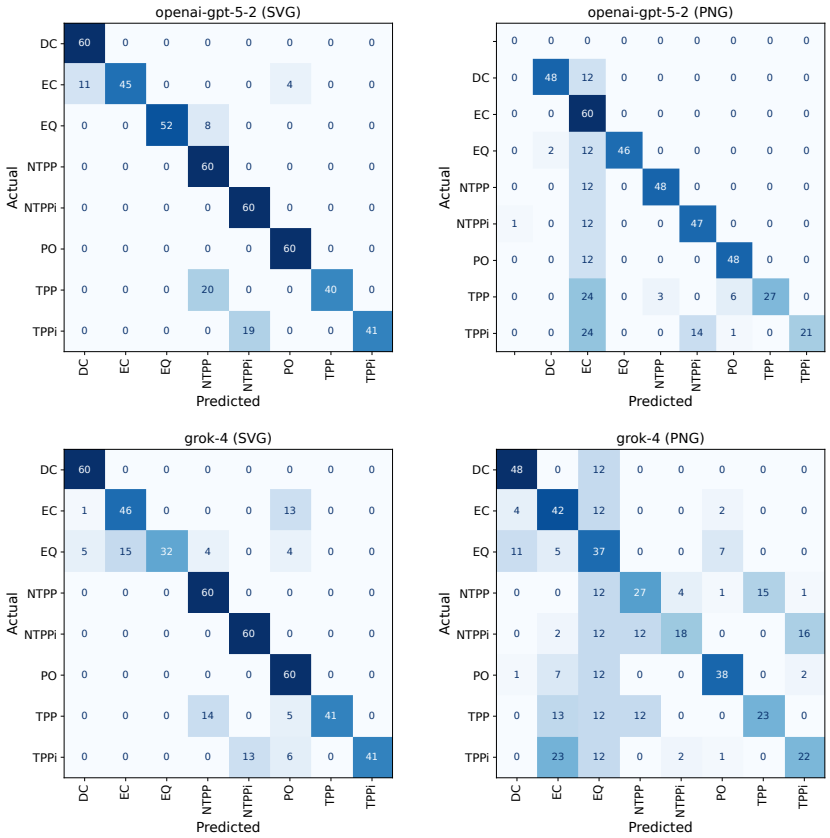

Figure 5 Relation diagram recognition confusion matrices for the best performing models (*GPT-5.2* and *Grok 4*), with SVG on the left and PNG on the right.

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.

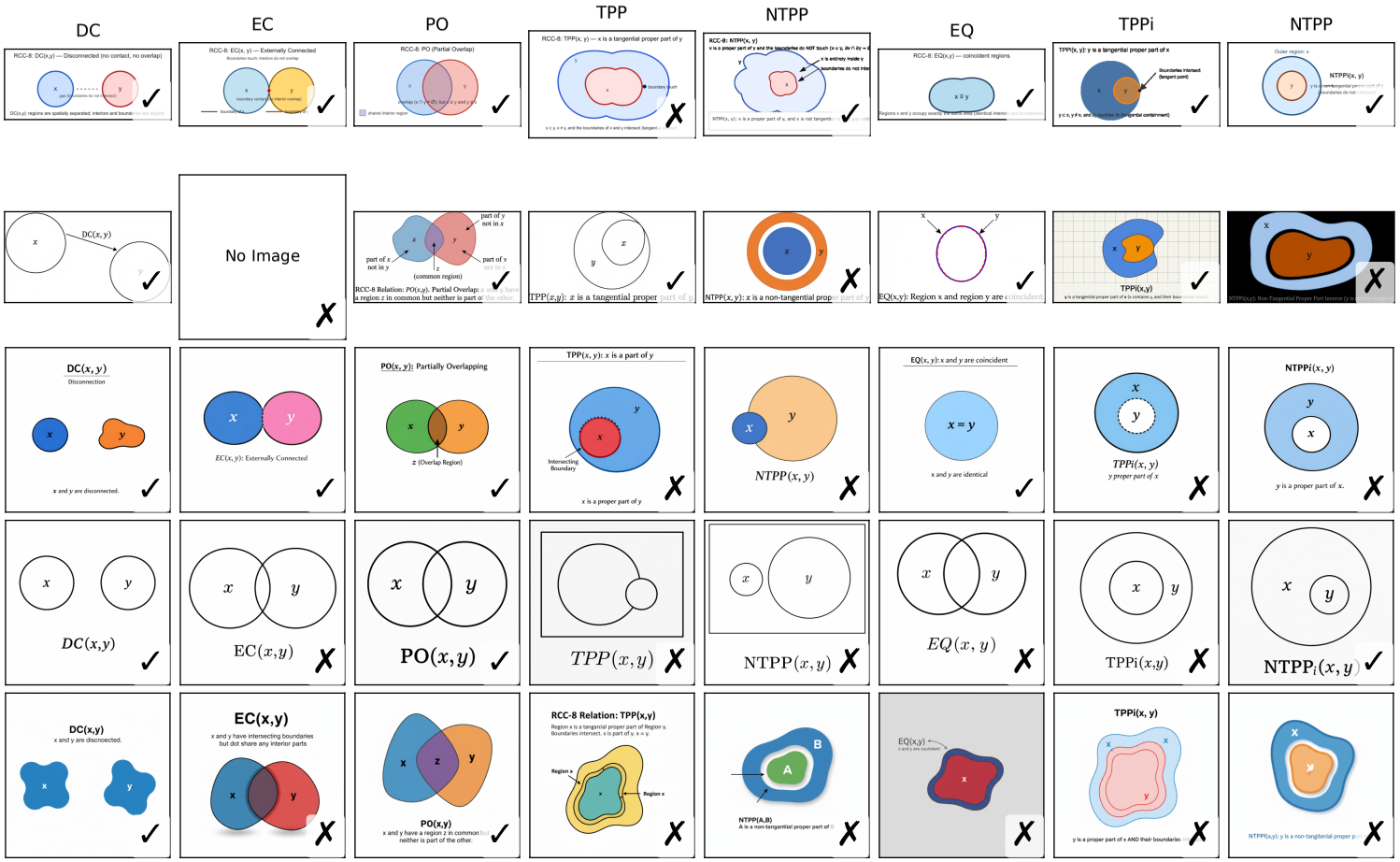

Figure 6 Relation diagram generation by model. gpt-5.2 (high) with SVG output (top, score 7), gemini-3-pro-image-preview (score=5), gpt-image-1-5 (score=4), gpt-image-1 (score=3), and gemini-2-5-flash-image (score=2). z-image-turbo scored 0 and is not shown. Ticks and crosses represented correct and 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.

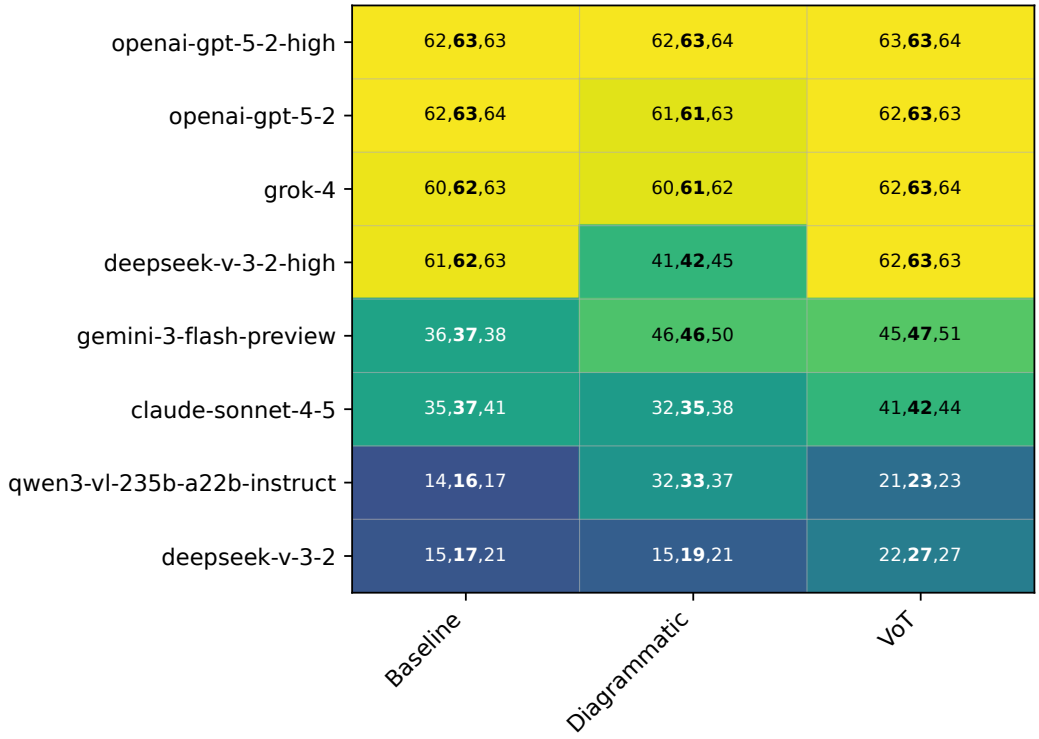

Figure 7 RCC-8 CT correct answers by model for baseline, diagrammatic reasoning and Vision of Thought experiments. Each model run has 64 questions (eight relations × eight relations), so each score is out of 64. n=3 repeats are shown with the median in bold. Rows are ordered by the sum of the maximum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

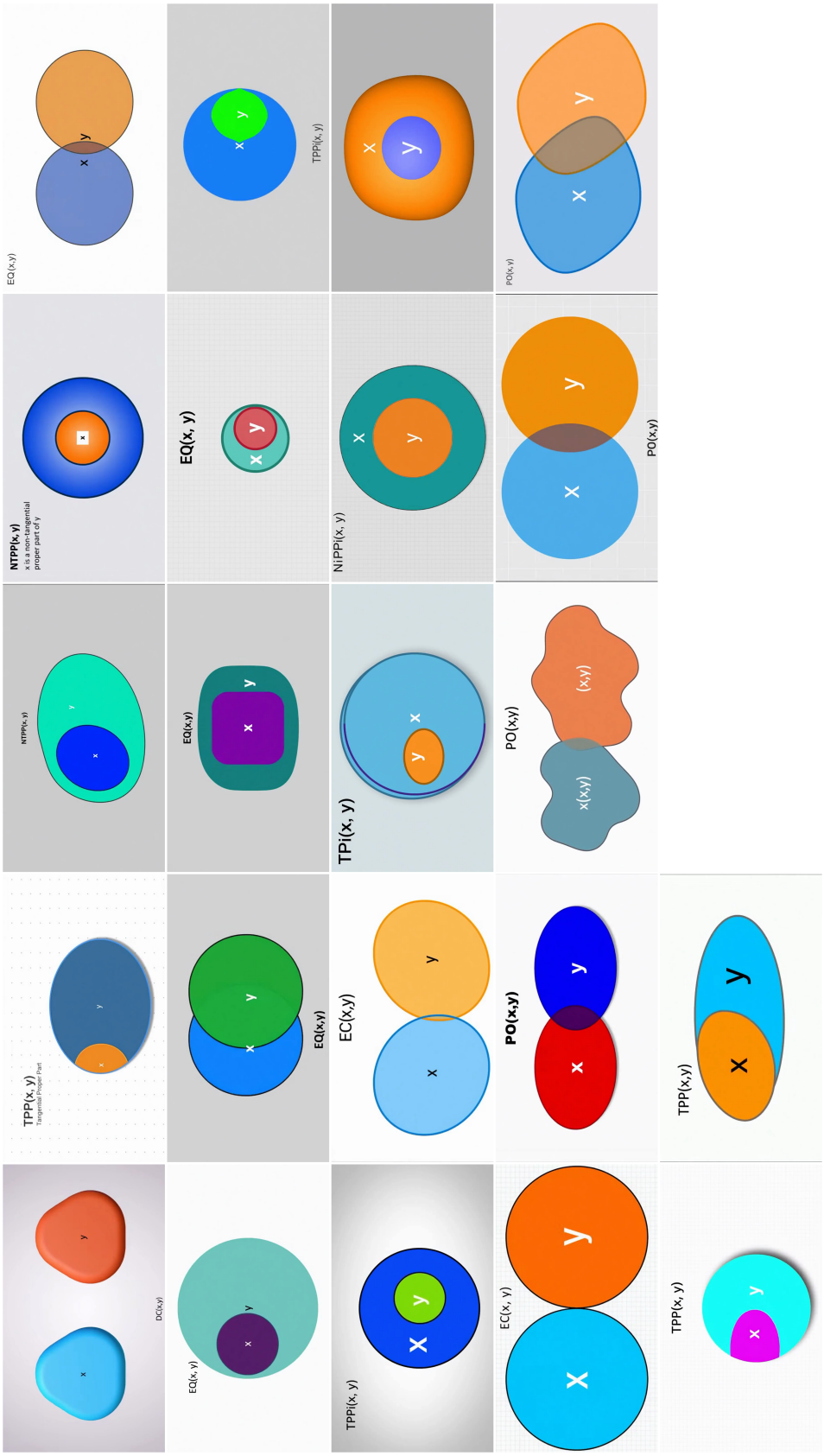

Figure 8 First frames from the *Sora 2* CN videos (n=1). This figure gives an overall impression of CN video generation results. The videos themselves are available in the GitHub repository that accompanies this paper.

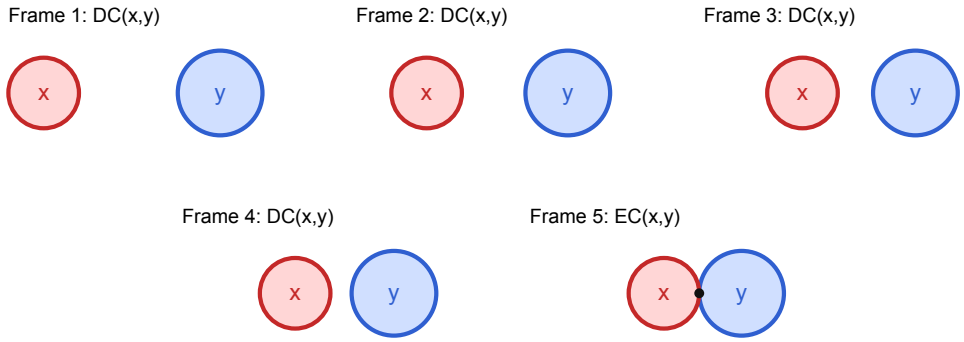

Figure 9 Frames generated by *GPT-5.2* for the CN transition from DC to EC (frame labels are part of the generated SVG). The generation also includes this commentary: Frame 1 – **DC(x,y)** (clearly separated) Frame 2 – **DC(x,y)** (closer, still a gap) Frame 3 – **DC(x,y)** (very close, still disconnected) 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.