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

Robert E Blackwell[1] , **Anthony G Cohn**[12]

[1]University of Leeds
[2]The Alan Turing Institute
{r.e.blackwell,a.g.cohn}@leeds.ac.uk

Abstract

Diagrams are central to qualitative spatial representation and reasoning, yet it remains unclear whether current multimodal foundation models can use them reliably. We evaluate four capabilities using RCC-8 as a benchmark: recognising relation diagrams, generating diagrams, answering composition questions, and illustrating conceptual-neighbourhood transitions. Results show that current models are promising but not yet dependable. Vision-language models recognise vector diagrams better than raster diagrams but struggle with relations requiring exact boundary contact. Text-to-image and video models often generate visually plausible but topologically incorrect diagrams. LLMs can solve many RCC-8 composition questions, but prompts to visualise or draw diagrams do not consistently improve performance.

1 Introduction

Diagrams often make spatial relations directly perceptible, and hence support reasoning in ways that purely sentential descriptions do not [Tylén *et al.*, 2014]. This is especially clear in qualitative spatial reasoning, where calculi are commonly defined using diagrams. The Region Connection Calculus RCC-8 [Randell *et al.*, 1992; Cohn *et al.*, 1997] is a well-known example consisting of eight binary topological relations between regular closed regions: DC (Disconnected), EC (Externally Connected), PO (Partially Overlapping), EQ (Equals), TPP (Tangential Proper Part), NTPP (Nontangential Proper Part), and the inverses TPPi and NTPPi. These relations support both a composition table, specifying possible x–z relations given x–y and y–z relations, and a conceptual neighbourhood (CN) graph, specifying which relations can be reached by continuous motion or deformation.

Recent foundation models extend textual large language models (LLMs) with multimodal capabilities. Vision-language models (VLMs) incorporate vision encoders to process visual information, while image and video generation models enable the synthesis of images and videos, respectively. This raises a natural question: can such models recognise, generate and exploit diagrams pertaining to qualitative spatial reasoning? We address this question using RCC-8 as

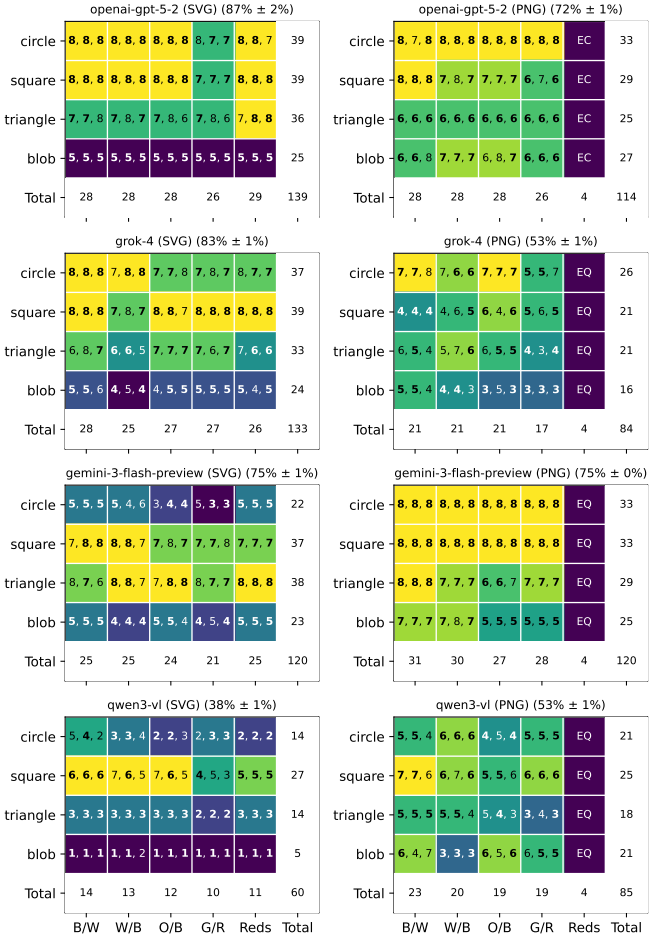

Figure 1: Vision-Language Model relation diagram recognition comparing shape and colour combinations for SVG and PNG by model. Each model run has 160 questions (8 relations × 5 colour combinations × 4 shapes), so the score in each cell is out of 8, with n=3 repeats. The colour combinations are b(lack)/w(hite), w(hite)/b(lack), g(reen)/r(ed), o(range)/b(lue) and adjacent reds. 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.

a benchmark. RCC-8 is small enough for exhaustive testing, but difficult enough to expose important failure modes: in particular, several distinctions depend on exact boundary contact, such as EC versus DC, or TPP versus NTPP.

This paper is a summary of a longer version [Blackwell and Cohn, 2026], where full details can be found.

2 Benchmark and Experimental Design

We evaluate four diagrammatic capabilities. First, we test whether VLMs can recognise diagrams depicting each RCC-8 relation. We generate diagrams for all eight relations using four classes of shape: circles, squares, triangles and irregular blobs, and five colour schemes. Each diagram is presented both as SVG and as PNG. The model is asked to answer with one of the eight RCC-8 relation names only.

Second, we test relation diagram generation. For each RCC-8 relation R, we give a short definition of RCC-8 and ask image-generation models to draw a diagram illustrating $R(x, y)$. We also test whether an LLM can generate SVG diagrams when explicitly instructed to do so.

Third, we test compositional reasoning. We use the 64 entries of the RCC-8 composition table, asking questions of the form: if $R_1(x, y)$ and $R_2(y, z)$ hold, what are all the possible RCC-8 relations that can hold between x and z? We compare a baseline prompt with two diagram-related prompts: a Vision-of-Thought style instruction to visualise intermediate states [Wu *et al.*, 2024], and an explicit instruction to draw labelled diagrams before answering.

Fourth, we test whether models can illustrate CN transitions. For each adjacent pair of RCC-8 relations, we ask a video model to generate a short animation, ask image generation models to generate five successive frames, and similarly ask an LLM to generate frames in SVG. Outputs are scored as correct only when the qualitative relation is represented correctly; visual plausibility alone is insufficient.

3 Results

In relation recognition, no VLM recognised all diagrams correctly (Figs 1, 2). The strongest models, *GPT-5.2* and *Grok 4*, performed substantially better on SVG than PNG: respectively, 87% versus 72%, and 83% versus 53%. Relations not requiring exact tangency, such as DC, PO, NTPP and NTPPi, were recognised most reliably. By contrast, tangential relations were a consistent source of errors. The most common confusions were TPP as NTPP, TPPi as NTPPi, and EC as DC. This suggests that models are still weak at detecting exact boundary contact, which is central to RCC-8. *GPT-5.2* was able to reliably recognise EC when relations were presented in PNG format, but the confusion matrix in Figure 3 shows that *GPT-5.2* tends to overestimate EC so this result is not surprising.

Shape and colour also mattered. Diagrams using circles or squares were recognised more accurately than diagrams using triangles or irregular blobs. Colour had little effect for SVG, where colour is simply an attribute in a vector representation. In PNG, however, black/white diagrams were generally easiest, orange/blue diagrams were more reliable than red/green diagrams, and two adjacent shades of red were

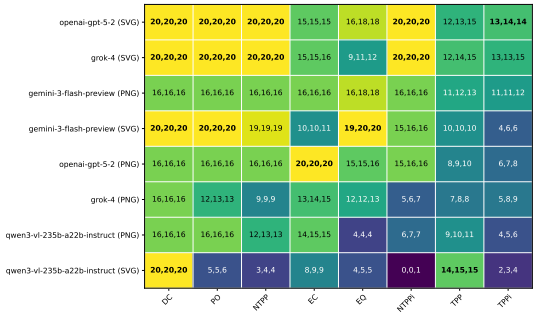

Figure 2: Vision-Language Model relation diagram recognition for each RCC-8 relation. Each model run has 160 questions (8 relations $\times$ 5 colour combinations $\times$ 4 shapes), and so each score is out of 20, with n=3 repeats. Rows and columns are ordered by the sum of the median scores, so models 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.

hardest. 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 general, this suggests that raster recognition remains sensitive to low-level perceptual cues such as contrast, edge salience and foreground/background separation.

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 3). On one occasion, *GPT-5.2* failed to give an answer to a PNG question at all, when the answer was NTPPi.

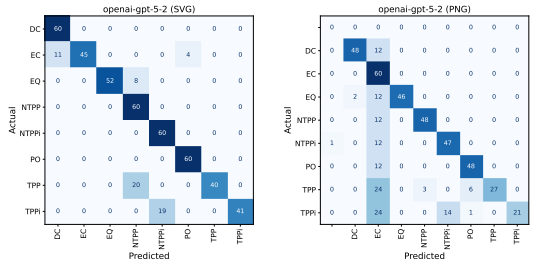

Figure 3: Relation diagram recognition confusion matrices for the best performing model (*GPT-5.2*), with SVG (left) and PNG (right).

For relation generation (Fig. 5) the strongest result came from *GPT-5.2* when asked to output SVG, scoring 7/8. Among image-generation models, *Gemini 3 Pro Image Preview* performed best, scoring 5/8. Other raster image generators produced visually plausible diagrams but often failed on exact RCC-8 distinctions. This is significant because RCC-8 correctness depends not merely on approximate appearance but on precise topology. For example, the tangential distinction between TPP and NTPP.

The strongest LLMs answered many RCC-8 composition questions correctly, with *GPT-5.2* achieving 64/64 in some

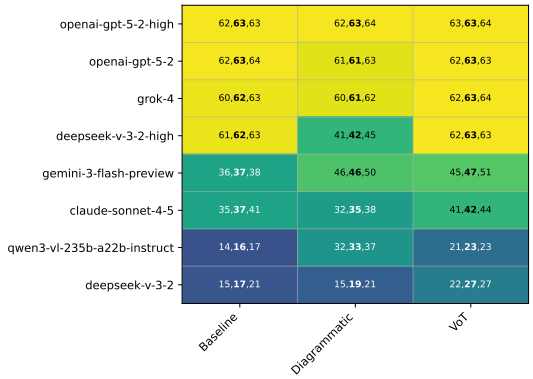

Figure 4: RCC-8 composition-table accuracy under baseline, diagrammatic-reasoning and Vision-of-Thought prompts.

repeats. However, explicit diagrammatic-reasoning prompts did not reliably improve the strongest models. Weaker models showed modest gains from Vision-of-Thought or diagrammatic prompts, but not consistently. Thus, improved performance under such prompts should not be taken as evidence of genuine diagrammatic reasoning: it may instead reflect a more general prompting effect, such as encouraging more careful step-by-step reasoning.

The conceptual-neighbourhood experiments were particularly challenging. *Sora 2* produced plausible-looking videos, but only 3/22 were judged plausible illustrations of the requested RCC-8 transition, and only 13/22 correctly depicted the starting relation in the first frame. By contrast, SVG frame generation by *GPT-5.2* produced plausible five-frame transitions for all 22 tested conceptual-neighbourhood transitions, although some outputs were cropped or stylistically inconsistent. This again suggests that explicit vector generation is currently more reliable than raster or video generation for topologically constrained diagrams.

4 Discussion and Conclusion

The experiments indicate that current multimodal foundation models have partial but fragile competence with diagrammatic qualitative spatial reasoning. Recognition is better for vector graphics than raster graphics, and better for simple, regular shapes than for irregular ones. The main failure mode is not gross spatial layout but precise topological distinction: boundary contact, containment with non-zero margin, and inverse containment relations remain difficult.

The generation results reinforce this conclusion. Image and video models can produce appealing visual outputs, but RCC-8 requires correctness under exact qualitative constraints. A diagram that is aesthetically plausible can still be wrong. SVG generation is more successful, plausibly because coordinates and shapes can encode topological constraints more explicitly than diffusion-style raster generation.

The composition results show that modern LLMs can perform surprisingly well on RCC-8 composition questions, but they do not show clear evidence of benefiting from explicit diagrammatic prompts. This is important because asking a model to "visualise" or "draw" does not by itself establish that the model is using an external diagrammatic representation in the human sense. The benchmark therefore distinguishes three abilities that might otherwise be conflated: recognising diagrams, generating diagrams, and using diagrams to improve reasoning.

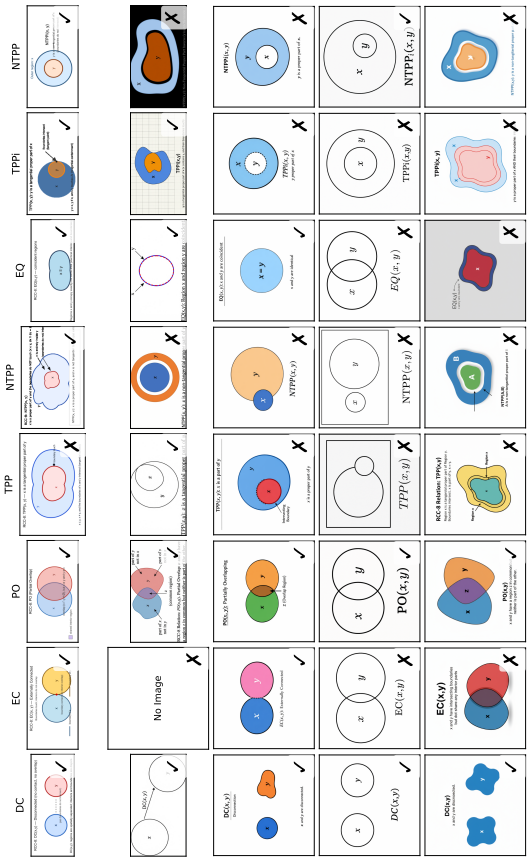

Figure 5: Relation diagram generation by model. gpt-5.2 (high) with SVG output (top, score 7), gemini-3-pro-image-preview (score=5), gpt-image-1-5 (score=4), gpt-image-1 (score=3), and gemini-2-5-flash-image (score=2). z-image-turbo scored 0 and is not shown. Ticks and crosses represented correct and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

RCC-8 is a compact but demanding benchmark for multimodal foundation models. It exposes failures that are easy to miss in ordinary image understanding tasks, because the relevant distinctions are qualitative and topological rather than merely visual. Future work could extend the benchmark to other calculi, automate scoring of generated diagrams, test additional graphical formats, and investigate whether models can be trained or constrained to produce diagrams that are not only plausible, but spatially correct.

Data Availability: All prompts, API requests, responses and experimental results are available in the accompanying repository[1], following recommendations for transparent reporting of LLM evaluations [Burnell *et al.*, 2023].

[1] github.com/RobBlackwell/cosit2026, last accessed July 2026

References

[Blackwell and Cohn, 2026] Robert E Blackwell and Anthony G Cohn. RCC-8 as a benchmark for diagrammatic reasoning in multimodal foundation models. In *Proc. COSIT, to appear*, 2026.

[Burnell *et al.*, 2023] 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.

[Cohn *et al.*, 1997] 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.

[Randell *et al.*, 1992] David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proc. KR*, pages 165–176, 1992.

[Tylén *et al.*, 2014] 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.

[Wu *et al.*, 2024] 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.