Introduction

This research project investigates the fundamental visual "blind spots" of Claude's Haiku 4.5 model, specifically testing performance on three core perceptual categories:

Localization: Identifying the precise coordinates and geometric anchor points of objects. This is foundational for tasks such as reading tables, charts, and dashboards, where small localization errors can propagate into incorrect numerical values or trend interpretations.

Label Association: Linking text labels to data points. This capability is critical for interpreting legends, annotations, and multi-series charts, and failures here often manifest as confident misattribution even when the underlying data is visually clear.

Path Following: Tracking the continuity of a line through intersections/occlusions. This primitive is central to understanding flowcharts, dependency graphs, and decision diagrams, where errors in visual continuity can lead to fundamentally incorrect structural interpretations.

Additional experiments (e.g., counting and relative magnitude) were conducted but not included in this report; details and code are provided in the accompanying repository.

Experiments

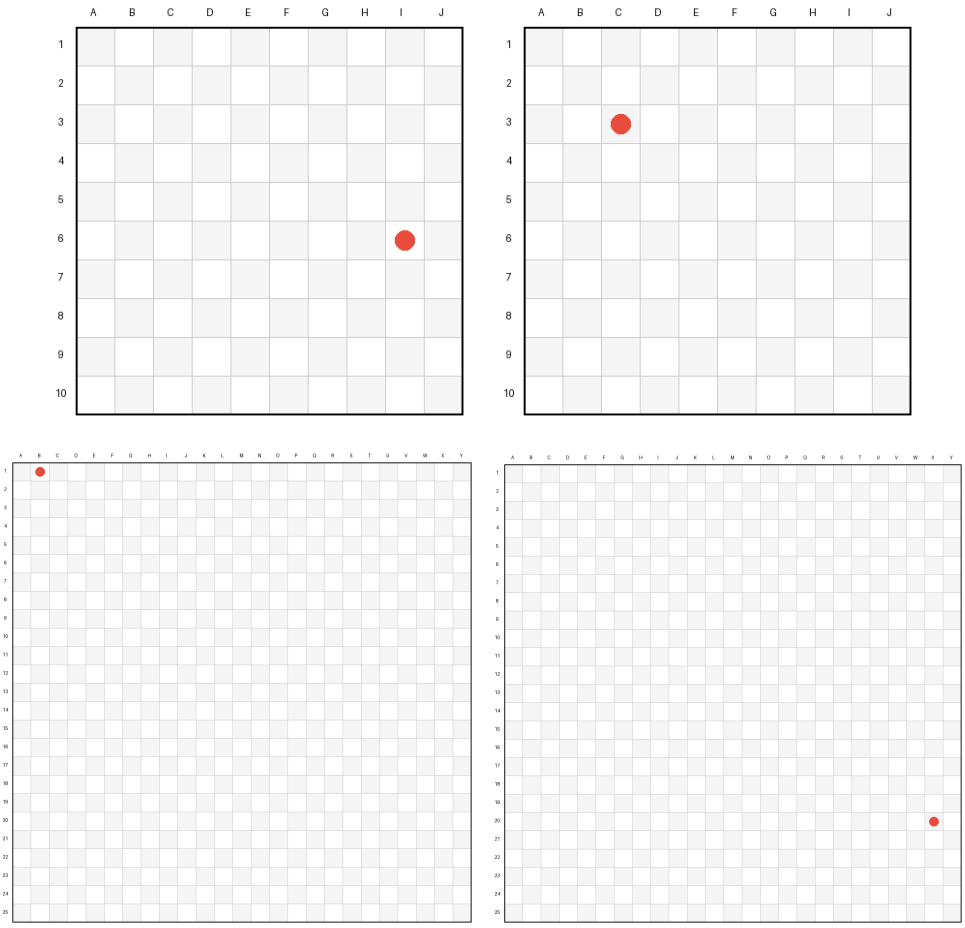
For each perceptual category, I evaluated two difficulty variants of the same task, both with and without thinking enabled (1024-token budget), resulting in four experiments per category.

Details for each experiment are provided in the next section of this report. All reported confidence intervals are 95% confidence intervals calculated using the Wilson score method.

Localization

Task

In this task, the model is asked to identify the grid cell containing a single red dot.



Results

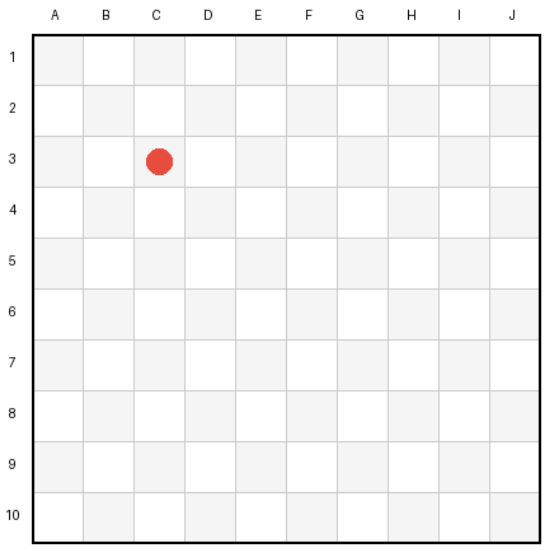
10x10 Grid:

Condition	Correct	Accuracy CI
No Thinking	945 / 1000	92.9% - 95.7%
With Thinking	973 / 1000	96.1% - 98.1%

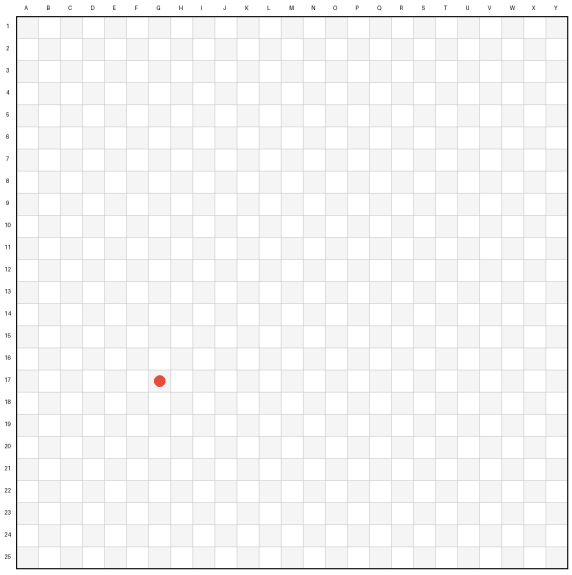
25x25 Grid:

Condition	Correct	Accuracy CI
No Thinking	487 / 1000	45.6% - 51.8%
With Thinking	404 / 1000	37.4% - 43.5%

Failure Cases



Groundtruth: C3; Prediction: D3



Groundtruth: G17; Prediction: F17

Discussion

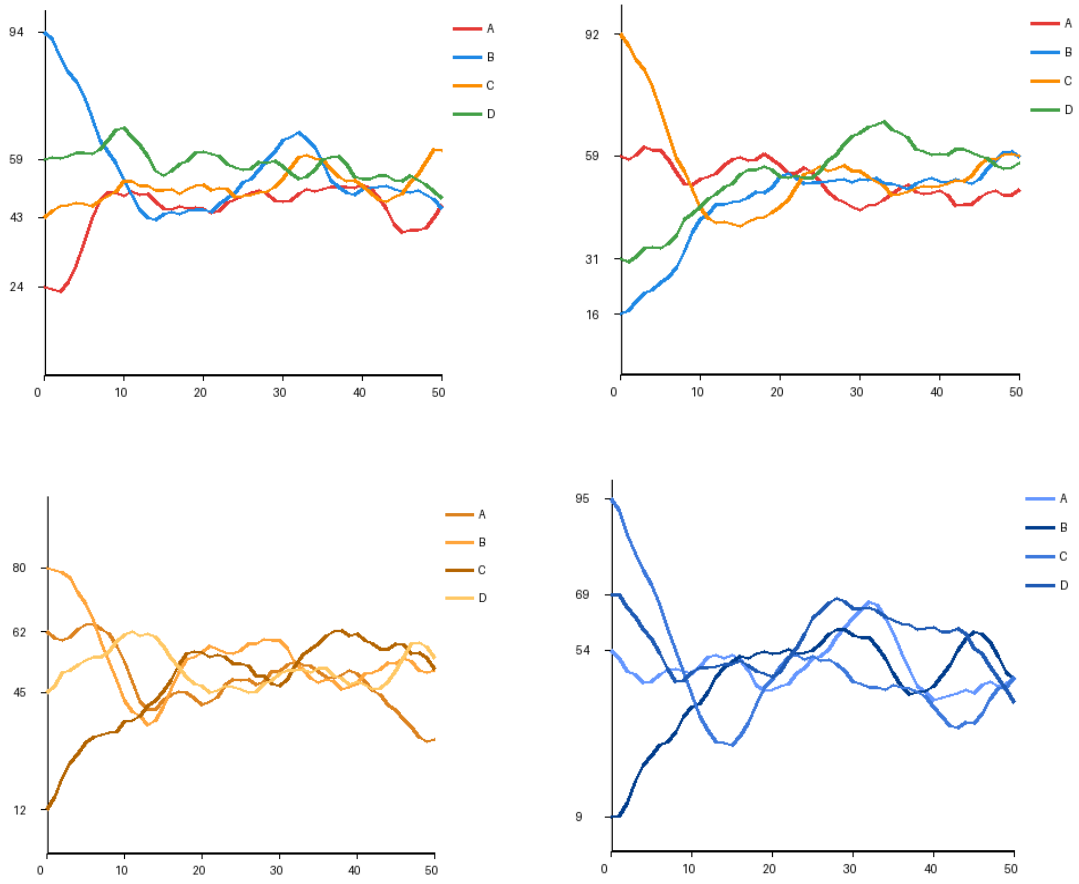
Grid retrieval accuracy degrades sharply with grid size (≈93–98% at 10×10 vs ≈37–52% at 25×25). Thinking slightly helps when spatial precision is preserved but degrades performance once localization collapses at the perceptual level, likely amplifying incorrect hypotheses.

Label Association

Task

Each image contains a line chart with 4 lines of 4 different colors, and a legend labeling each line. The prompt asks the value of one of the lines at x=0, given the label of a line. We perform 2 variations of this task, one with 4 distinct colors, and one with 4 shades of the same color.

Example prompt: “What is the value of line A at x=0? Choose from one of the following options: 90, 66, 30, 21”



Results

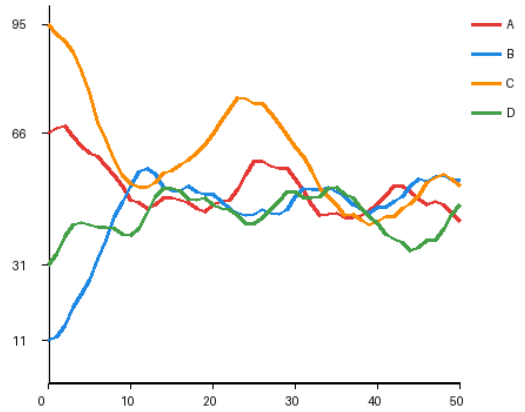
Distinct Colors:

Condition	Correct	Accuracy CI
No Thinking	968 / 1000	95.5% - 97.7%
With Thinking	984 / 1000	97.4% - 99.0%

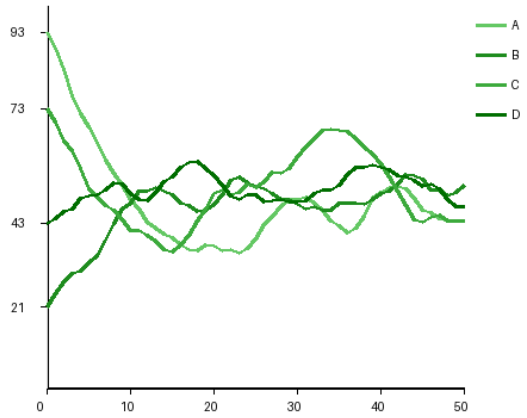
Same Color, Different Shades:

Condition	Correct	Accuracy CI
No Thinking	460 / 1000	42.9% - 49.1%
With Thinking	397 / 1000	36.7% - 42.8%

Failure Cases



Line: A; Groundtruth: 66; Prediction: 95



Line: B; Groundtruth: 21; Prediction: 43

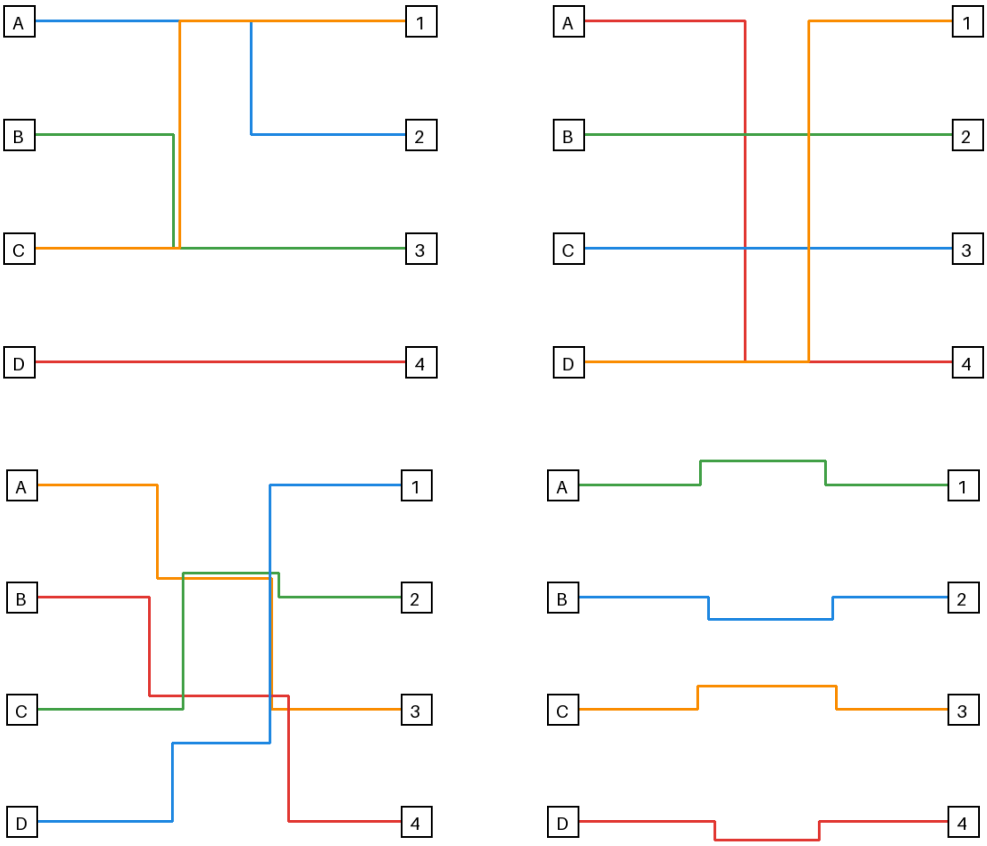
Discussion

For the line chart task, accuracy is high with distinct colors (96–99%) but drops sharply under low chromatic contrast (37–49%). Thinking provides no benefit in the easy regime and slightly degrades performance in the hard regime, indicating that the failure mode is unreliable visual discrimination rather than legend parsing or numerical reasoning.

Path Following

Task

In this, the model is asked to identify the numbered box that is connected to a given letter box.



Results

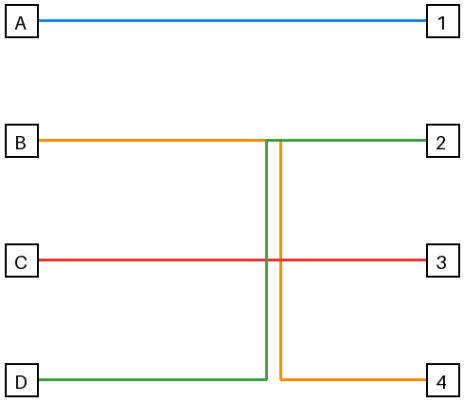
2 Bends:

Condition	Correct	Accuracy CI
No Thinking	420 / 1000	39.0% - 45.1%
With Thinking	654 / 1000	62.4% - 68.3%

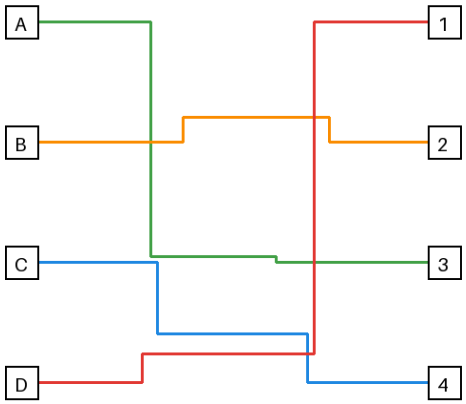
4 Bends:

Condition	Correct	Accuracy CI
No Thinking	459 / 1000	42.8% - 49.0%
With Thinking	893 / 1000	87.2% - 91.1%

Failure Cases



Box D; Groundtruth: 2; Prediction: 4



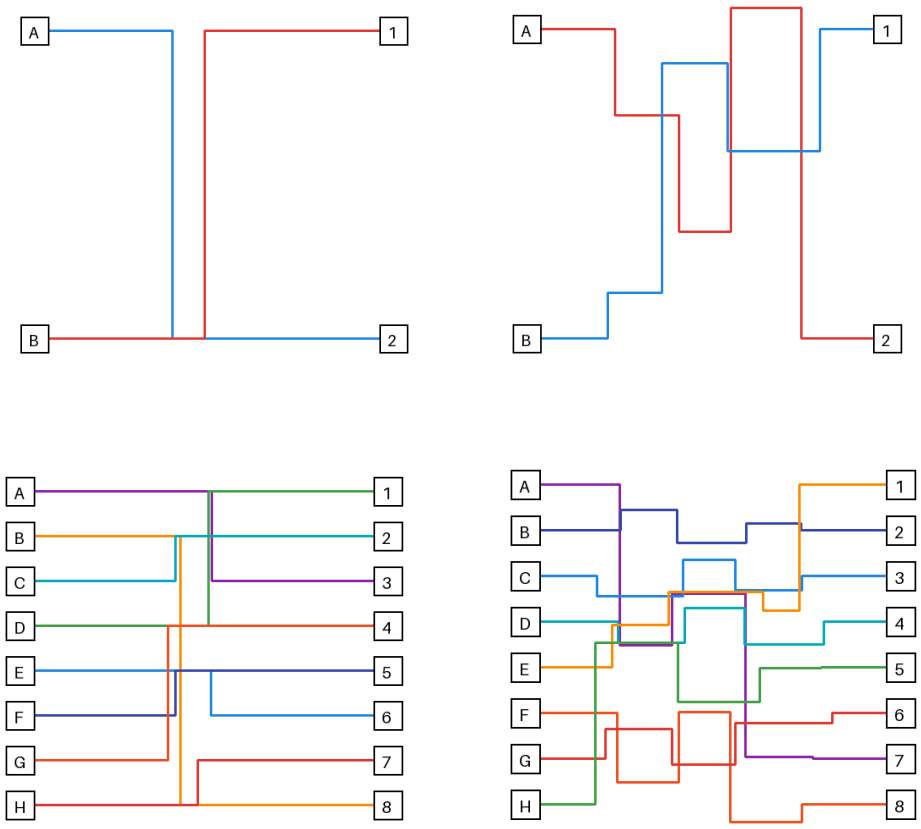
Box D; Groundtruth: 1; Prediction: 4

Discussion

Thinking substantially improves path-following accuracy (42.0% → 65.4% with 2 bends; 45.9% → 89.3% with 4 bends), yet 2-bend cases are consistently harder than 4-bend cases. Inspection shows this is due to increased path overlap in low-bend layouts, causing systematic path-identity switching despite distinct colors.

Deep Dive: Path Following

To further analyze path following, I evaluated accuracy across combinations of path bends and number of boxes, both with and without thinking enabled.



	2 bends	**4 bends**	**6 bends**	**8 bends**
2 boxes	73.9% / 83.0%	74.7% / 97.6%	73.8% / 98.5%	72.4% / 97.8%
4 boxes	43.0% / 65.4%	45.9% / 89.3%	49.6% / 92.6%	48.0% / 92.4%
6 boxes	23.3% / 44.4%	34.5% / 72.1%	32.3% / 78.0%	35.8% / 81.4%
8 boxes	19.8% / 35.1%	26.0% / 50.7%	25.3% / 55.3%	28.6% / 59.6%

Note: Each cell in the table corresponds to the accuracy without thinking enabled (on the left) and with thinking enabled (on the right).

Evaluating path-following accuracy across bend counts and number of boxes reveals three patterns: (1) thinking consistently improves accuracy, (2) performance is largely insensitive to bend count except where low-bend layouts induce path overlap, and (3) increasing the number of boxes sharply degrades accuracy due to path-identity ambiguity.

Inspection of model reasoning traces reveals explicit hypothesis revision (e.g., “Actually, let me look more carefully at the connections”), suggesting that thinking helps preserve path identity under temporary visual ambiguity.

Based on these observations, I ran two follow-up experiments: (1) varying the thinking token budget while holding task complexity fixed, and (2) re-running the 2-bend condition with prompts that explicitly warn about overlapping paths and emphasize color-based identity cues.

Varying Thinking Budget

In this experiment, we fix the number of boxes to be 4, and the number of bends to be 4, and we vary the thinking budget.

Thinking Tokens Budget	**1024**	**1500**	**2000**	**5000**
Accuracy CI	87.2% - 91.1%	87.1% - 91.0%	88.1% - 91.8%	90.6% - 93.9%

Increasing the thinking budget beyond 1024 tokens does not produce statistically significant gains, indicating that reasoning depth is not the limiting factor once a viable path hypothesis is formed.

Better Prompting

For this experiment, I varied the number of boxes, fixed the number of bends to be 2, fixed the thinking budget to 1024 tokens, and updated the prompt to be clear about using color as a cue and how to handle overlapping lines. The updated prompt looked like this (diff in red):

“Follow the path starting from box A. The path will only be one continuous color, but other paths may overlap with it, obscuring part of the path. Be sure to follow the path with the correct color, and determine the answer to this question; Which numbered box does it lead to?”

Number of Boxes	**Previous Prompt CI**	**Updated Prompt CI**
2	80.5% - 85.2%	87.7% - 91.4%
4	62.4% - 68.3%	71.2% - 76.6%
6	41.3% - 47.5%	54.2% - 60.3%

Unsurprisingly, better prompting that addresses Claude’s weaknesses has a significant positive impact on accuracy.

Finetuning Proposal

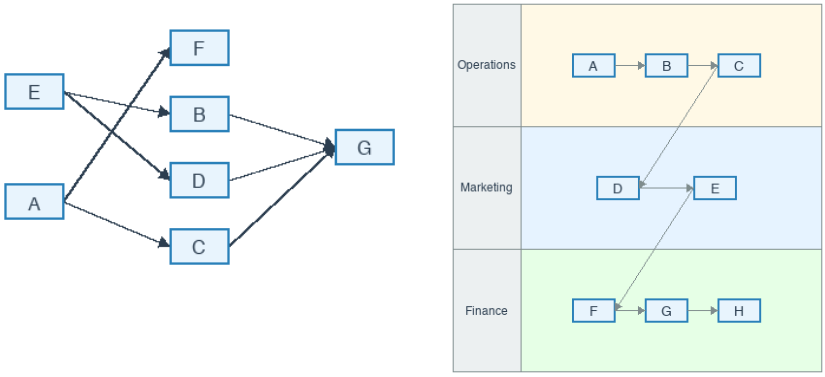
Given that path-following is a verifiable task with procedurally generated ground truth data, if I were at Anthropic, I'd start by understanding how the team does finetuning for visual tasks and verifiable tasks like math and coding. I'd also look for existing best practices for synthetic data generation. The following is based on my own intuition and online research.

Data

I would generate a large synthetic dataset focused specifically on path-following primitives, where the core task is to identify and preserve the identity of a visual path as it connects two endpoints through intersections, overlaps, and occlusions. This would include cases such as:

- 2-bend paths that overlap without crossing
- Paths that cross each other at a 4-cross junction
- Varying number of paths in a single image

I would also generate a mix of simple synthetic data, like used in this study, as well as more realistic synthetic data. The realistic synthetic data would mirror real-life business use cases more closely. This would include charts/diagrams such as Sankey diagrams, flowcharts, logic trees, circuit diagrams, decision trees, etc. To ensure a high degree of realism, I would incorporate things like varying fonts and font sizes, realistic layout conventions, visual clutter, and partial occlusion.



Training

I would employ a mix of supervised finetuning and reinforcement learning. I would run SFT first to teach stable visual parsing and consistent structured outputs, then RL to calibrate the model's decision policy in difficult/high-ambiguity examples.

Supervised Finetuning

I would train on a 50/30/20 split of simple/moderately perturbed/hard cases, balanced across task types, with outputs constrained to a strict schema. The simple data teaches the core visual

operator, while the hard/realistic data teaches robustness to layout conventions, text density, and clutter. I would start training with a higher percentage weighted towards the simple data, then slowly increase the percentage of realistic data.

Reinforcement Learning

For RL, I would concentrate on the failure slice: 60–70% overlap/adjacency/high-complexity cases, 20–30% crossings/near-miss attachments, remainder clean for retention, gradually increasing difficulty via a curriculum (no overlap → overlap stems → long overlaps and near-parallel → occlusion/compression and distractors).

I would use deterministic exact-match rewards (1.0 for correct answer, 0.0 otherwise). I would also experiment with:

- Partial rewards via partial path correctness
- A controlled subset of ambiguous examples with an allowed “unknown” response, rewarding calibrated abstention over confident misclassification

Food for Thought

One possible contributing factor to several of these failures may lie upstream of finetuning, in image tokenization and patch-level representation. Both the sharp degradation in grid localization at higher resolutions and the identity-switching failures under overlapping paths are consistent with information loss at the token level, where multiple visual elements collapse into similar embeddings. While targeted finetuning can likely mitigate these effects, it may be valuable to study whether higher-resolution or adaptive tokenization further improves robustness on these perceptual primitives.