

Chain-of-Thought Degrades Visual Spatial Reasoning Capabilities of Multimodal LLMs

Sai Srinivas Kancheti^{1,*†}, Aditya Sanjiv Kanade^{2,*}
Vineeth N. Balasubramanian^{1,2}, Tanuja Ganu²

¹Indian Institute of Technology Hyderabad, India ²Microsoft Research India
cs21resch01004@iith.ac.in, kanade850@gmail.com
{vineeth.nb, tanuja.ganu}@microsoft.com

Abstract

Multimodal Reasoning Models (MRMs) leveraging Chain-of-Thought (CoT) based thinking have revolutionized mathematical and logical problem-solving. However, we show that this paradigm struggles with generalized spatial intelligence. We perform a comprehensive evaluation of twenty-one models across thirteen spatial benchmarks and identify a critical gap: CoT prompting consistently degrades performance in visual spatial reasoning. Furthermore, through a novel *No-Image++* ablation, we demonstrate that MRMs and CoT prompted MLMs suffer from severe shortcut learning, and hallucinate visual details from textual priors even when the image is absent. These findings challenge the efficacy of text-only CoT for spatial tasks and underscore the need for vision-centric reasoning paradigms¹.

1 Introduction

The emergence of "System 2" Multimodal Reasoning Models (MRMs), models post-trained via SFT and RL to generate step-by-step reasoning, has driven remarkable progress in mathematical and logical domains. By leveraging Reinforcement Learning (RL) (Lambert et al., 2024; Guo et al., 2025) and long Chain-of-Thought (CoT) (Wei et al., 2022; Wang et al., 2022) inference, MRMs demonstrate the ability to self-correct and reason through complex problems. Separately, *CoT prompting* is a general technique that instructs any Multimodal Language Model (MLM) to think step-by-step before answering. However, a fundamental question remains: **does this text-centric reasoning paradigm translate to spatial intelligence?** Spatial reasoning requires grounding, geometric intuition, and precise localization, which are skills

that may not easily arise from verbose, text-based reasoning (Tong et al., 2024a,b).

In this work, we conduct a comprehensive evaluation of twenty-one models: eight state-of-the-art open-source MRMs (e.g. GThinker, ViGoRL, TreeVGR), eight diverse backbone MLMs spanning three families (Qwen, InternVL, LLaVA), and five proprietary models (GPT-4o, GPT-4.1-mini, GPT-5, GPT-5-mini, GPT-5-nano). We benchmark these models across thirteen datasets covering static 2D relations, 3D geometry, and dynamic/temporal understanding. To isolate the impact of CoT reasoning, we standardize our evaluation using a uniform evaluation and scoring policy. Our findings reveal that contrary to trends in other domains, CoT prompting degrades performance in visual spatial tasks. Our contributions are as follows: (i) We show that MRMs consistently underperform their own backbone on generalized spatial benchmarks. In our experiments, 7 out of 8 reasoning models failed to surpass the backbone they were distilled from. (ii) We demonstrate in Figure 1 that CoT prompting lowers accuracy by an average of 3% across a diverse range of MLMs. (iii) Through a novel *No-Image++* ablation, we show that MRMs suffer from severe shortcut learning. When presented with a blank image and a "Cannot determine" option, reasoning models continue to hallucinate visual details and confidently select incorrect answers based solely on textual priors.

These results suggest that simply scaling text-based reasoning is insufficient for robust spatial intelligence, highlighting the need for vision-centric training paradigms.

2 Methodology

We begin by describing the baseline models we consider, datasets and our evaluation scheme.

Baselines: We rigorously benchmark the performance of twenty-one models:

*Equal contribution.

†Work done while at Microsoft Research India.

¹Code and Model Evaluations are available at: https://github.com/microsoft/CoT_degrades_VSR

| Model | CoT | Non-CoT |
|-----------|-------|-----------------|
| GThinker | 62.52 | 39.38 (-23.14%) |
| R1-Ov | 46.88 | 47.84 (+0.96%) |
| ViGoRL | 60.68 | 62.52 (+1.84%) |
| VL-Re. | 60.99 | 62.18 (+1.19%) |
| Vision-G1 | 63.26 | 62.85 (-0.41%) |
| Vision-R1 | 58.86 | 59.6 (+0.74%) |
| TreeVGR | 61.11 | 62.6 (+1.49%) |
| ThinkLite | 62.61 | 62.74 (+0.13%) |

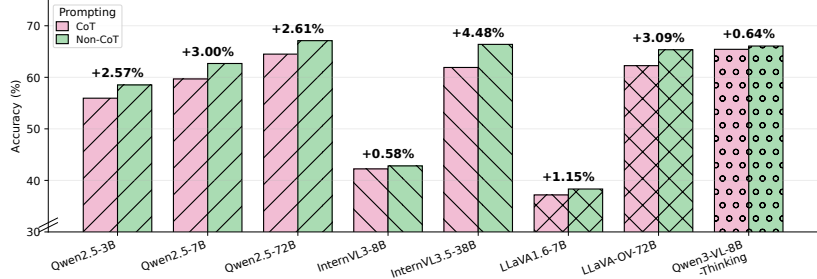


Figure 1: (Left) CoT vs Non-CoT performance of open-source MRMs. (Right) Bar chart showing the average accuracy of various families of MLMs over 13 benchmark datasets. For each model, the left bar shows the accuracy achieved by CoT prompting, and the right bar shows for base prompt (non-CoT). We observe CoT prompting drops performance over a wide range of backbones and model scales, including Qwen3-VL-8B-Thinking (Bai et al., 2025a), a model with explicitly enhanced spatial perception.

i) QwenVL backbone: We evaluate three scales of Qwen2.5-VL-Instruct (3B, 7B, 72B) (Bai et al., 2025b), along with Qwen3-VL-8B-Thinking (Bai et al., 2025a), a recent model with explicitly enhanced spatial perception post-trained to reason within `<think>` tags. We further evaluate eight SOTA general-purpose visual Multimodal Reasoning Models (MRMs) trained atop Qwen2.5-VL-7B using RLVR (Shao et al., 2024): GThinker-7B (Zhan et al., 2025) (Jun’25), ViGoRL-7B-Spatial (Sarch et al., 2025) (May’25), Vision-G1-7B (Zha et al., 2025) (Aug’25), R1-Onevision-7B (Yang et al., 2025b) (Mar’25), VL-Rethinker-7B (Wang et al., 2025b) (May’25), Vision-R1 (Huang et al., 2025) (Mar’25), TreeVGR (Wang et al., 2025a) (Jul’25), and ThinkLite-7B (Wang et al., 2025d) (Apr’25); **ii) InternVL backbone:** InternVL3-8B, InternVL3.5-38B (Wang et al., 2025c); **iii) LLaVA backbone:** LLaVA-v1.6-Mistral-7B, LLaVA-OneVision-Qwen2-72B (Li et al., 2024); and **iv) Proprietary frontier models:** GPT-4o (Hurst et al., 2024), GPT-4.1-mini (OpenAI, 2025), GPT-5, GPT-5-mini, and GPT-5-nano (Singh et al., 2025). In total, *twenty-one* diverse baselines cover reasoning and non-reasoning models, scales from 3B to 72B, and open-source to proprietary.

Open-source MRMs. We choose eight diverse top performing open-source MRMs (listed above) that are trained for general visual reasoning including spatial tasks. ViGoRL-spatial and TreeVGR are explicitly trained to perform spatial reasoning, while the remaining MRMs (with the exception of Vision-R1) contain spatial domains in their training data and are trained as general purpose visual reasoners. All eight MRMs are built atop the Qwen2.5-VL-7B-Instruct backbone, enabling a controlled

comparison against a shared backbone. A key motivation for our study, highlighted in Tab. 1, is that the original papers for these models evaluate on *math heavy* datasets which are *not vision-centric*. We hence comprehensively evaluate the reasoning capabilities of these models on spatial tasks.

| Baseline | Paper-Reported Datasets |
|-------------------|--|
| GThinker-7B | MMStar, RWQA, MMMU-Pro |
| R1-Onevision-7B | MathVision, Mathvista, Mathverse |
| ViGoRL-7B-Spatial | SAT-Val, BLINK |
| VL-Rethinker | MathVision, MMMU-Pro, MEGA |
| Vision-G1 | MathVista, MMMU-Pro, MMStar, ChartQA |
| Vision-R1 | MathVista, MMStar, ChartQA, MME _{sum} |

Table 1: Most eval datasets are Math heavy, are not vision-centric, and do not cover many aspects of spatial reasoning.

Datasets: We evaluate on thirteen datasets covering various aspects of spatial reasoning, which can be broadly categorized into static 2D datasets and 3D/dynamic datasets. The former datasets are usually confined to single images and focus on planar spatial relationships, usually from the camera’s perspective. We place BLINK (Fu et al., 2025), CV-Bench2D (Tong et al., 2024a), MMVP (Tong et al., 2024b), RealWorldQA (xAI, 2025), Spatial-Bench (Cai et al., 2025), VSR (Liu et al., 2023), and V*Bench (Wu and Xie, 2024) in this category.

We also consider benchmarks that require reasoning involving 3D geometry, depth, multi-image consistency, and temporal reasoning. These datasets often involve understanding the 3D position and relative orientation from the object’s perspective inside the image. 3DSRBench (Ma et al., 2024), CV-Bench3D (Tong et al., 2024a), MindCube (Yin et al., 2025), MMSIBench (Yang et al., 2025a), OmniSpatial (Jia et al., 2025), and SAT-Real (Ray et al., 2025) belong to this category.

We choose these datasets as they have real-world objects set in natural scenes, are difficult to answer on account of being vision-centric (Tong et al., 2024a), and cover a wide range of spatial capabilities. A summary of the datasets along with the various spatial facets they test is provided in Appx. Tab. 5. The left half of the table describes the 2D datasets and the right half the rest.

Evaluation: To ensure uniform evaluation, we follow VLMEvalKit (Duan et al., 2024) to provide a uniform system prompt as well as a uniform question format. All benchmarks are multiple-choice questions (MCQs) with options provided in the question prompt. The question format for all datasets is: Question:<question>\nOptions:\nA.<optA>\nB.<optionB> ... \nPlease select the correct answer (letter and option text) from the options above. We append dataset-specific prompts for OmniSpatial & MindCube to ensure good performance. The prompts are detailed in Appx. § A.

Metric. We use vLLM (Kwon et al., 2023) version 0.10.0 for performant, batched inference of MLMs on 4 NVIDIA A100 GPUs. We use a batch size of 16, set max new tokens generated to be 32768, set model context length as 32768, and perform inference on *bfloat16* precision. Proprietary models are accessed via the OpenAI ChatComplete API with the default temperature. We use pass@1 accuracy under greedy decoding (with temperature set to 0) as our metric. Open-source model results are over 3 seeds.

System Prompts. We evaluate models in two settings, using a i) *base prompt/non-CoT prompt* such as “You are a spatial-reasoning assistant. The user asks a question, and the Assistant solves it.” or a ii) *CoT prompt* where we append “First output the thinking process in <think></think> tags and then output the final answer in <answer></answer> tags.” to the base prompt. For CoT evaluation of MRMs, we use the custom CoT prompt they train on (instead of default shown above) for best performance as shown in the table below.

| Model | Custom CoT | Simple CoT |
|-------------|------------|------------|
| GThinker-7B | 62.52 | 59.57 |
| Vision-G1 | 63.26 | 62.06 |

The custom CoT prompt used for each MRM is shown in Appx § A.

Scoring the generations. We use an LLM-as-

a-judge along with a carefully designed prompt, shown in Appx. § A.3, to score all generations. We pick Qwen3-30B-A3B-Instruct-2507, a small non-reasoning text model as our judge for scoring since our evaluation is on MCQs with short answers i.e the final answer of the model is not free-form but restricted to the options provided. To validate our choice, we re-score the generations of Vision-G1 using GPT-4o as the judge, and compare it with our chosen judge. We observe a Cohen’s kappa score (Cohen, 1960) of > 0.99 indicating near-perfect agreement of judges.

3 Results and Analysis

(i) CoT Prompting Hurts Visual Spatial Reasoning. Contrary to trends in math and logic domains, we observe that CoT prompting frequently hurts performance in visual spatial tasks. Figure 1 (left) shows the performance of open-source MRMs under both CoT and non-CoT prompting. Surprisingly, these models, which have been explicitly trained via RL to reason, often perform better when this reasoning capability is suppressed. Six of the eight MRMs achieve higher accuracy with the non-CoT prompt than with their native CoT prompts. GThinker is an outlier, dropping -23.14% as it is not robust to changes in its prompt and generates ill-formed CoT-style traces even when given the direct-answer base prompt (example in Appx. § C) Figure 1 (right) shows that this trend holds across three families of models (Qwen, InternVL & LLaVA), and across a range of model strengths (params ranging from 3B to 72B). It also holds for Qwen3-VL-8B-Thinking (Bai et al., 2025a), a recent Qwen3-family model with explicitly enhanced spatial perception and a native <think>-mode where non-CoT still outperforms CoT by 0.64% at a strong baseline of $\sim 65\%$ (dataset-wise results in Appx. Table 8).

The trend further extends to proprietary models: Table 4 reports five GPT-family models, and non-CoT remains competitive with or exceeds CoT across the board. GPT-5 and GPT-5-nano in particular show clear CoT degradation ($+0.65\%$ and $+1.23\%$ for Non-CoT, respectively), mirroring the open-source trend, while GPT-4o and GPT-4.1-mini show only marginal CoT gains ($< 0.5\%$) despite the additional inference compute required for reasoning. We further analyze the CoT traces of proprietary models and find two notable differences from open-source models: (i) proprietary

| Models | 3DSRBench | BLINK | CV-Bench | | MindCube | MMSIBench | MMVP |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| | | | 2D | 3D | | | |
| Qwen2.5-VL-7B _{cot} | 57.11 _{0.39} | 53.44 _{0.30} | 75.92 _{0.03} | 76.09 _{0.31} | 30.83 _{0.25} | 27.47 _{0.21} | 72.44 _{0.68} |
| Qwen2.5-VL-7B | 55.38 _{0.06} | 56.04 _{0.03} | <u>77.17</u> _{0.03} | <u>83.78</u> _{0.04} | 35.11 _{0.18} | 26.87 _{0.09} | <u>75.78</u> _{0.32} |
| GThinker-7B | 56.58 _{0.20} | 54.76 _{0.17} | 77.40 _{0.06} | 82.95 _{0.04} | <u>40.16</u> _{0.32} | <u>27.33</u> _{0.31} | 73.78 _{0.42} |
| R1-Onevision-7B | 48.52 _{0.20} | 43.27 _{0.50} | 53.31 _{0.04} | 58.00 _{0.42} | 27.09 _{0.42} | 13.30 _{0.30} | 56.16 _{0.16} |
| ViGoRL-7B-Spatial | 55.84 _{0.20} | 52.51 _{0.26} | 76.59 _{0.27} | 86.14 _{0.10} | 39.36 _{0.16} | 25.87 _{0.17} | 73.22 _{0.42} |
| VL-Rethinker-7B | <u>56.99</u> _{0.11} | <u>54.60</u> _{0.23} | 76.06 _{0.12} | 80.75 _{0.14} | 37.81 _{0.27} | 26.90 _{0.08} | 75.89 _{0.16} |
| Vision-G1 | 55.91 _{0.01} | <u>54.60</u> _{0.08} | 76.70 _{0.15} | <u>83.75</u> _{0.14} | 38.10 _{0.31} | 26.07 _{0.12} | 76.56 _{0.16} |
| Vision-R1-7B | 55.01 _{0.20} | 46.47 _{0.66} | 71.58 _{0.12} | 75.83 _{0.31} | 36.95 _{0.59} | 22.90 _{0.36} | 72.22 _{0.32} |
| TreeVGR-7B | 51.53 _{0.03} | 53.16 _{0.25} | 76.24 _{0.09} | 75.17 _{0.13} | 44.25 _{0.66} | 27.17 _{0.33} | 71.33 _{0.27} |
| ThinkLite-7B | 57.26 _{0.13} | 57.13 _{0.08} | 76.89 _{0.18} | 80.44 _{0.16} | 30.13 _{0.23} | 27.97 _{0.56} | 73.56 _{0.16} |

| Models | OmniSpatial | RealWorldQA | SAT | SpatialBench | VSR | V*Bench | Avg. |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|--------------|
| Qwen2.5-VL-7B _{cot} | 40.40 _{0.80} | 63.05 _{0.27} | 59.22 _{0.57} | 61.75 _{0.70} | 81.83 _{0.35} | 76.27 _{0.25} | 59.68 |
| Qwen2.5-VL-7B | 45.23 _{0.11} | <u>69.02</u> _{0.00} | <u>63.11</u> _{0.16} | <u>62.87</u> _{0.00} | <u>85.38</u> _{0.04} | 79.06 _{0.00} | <u>62.68</u> |
| GThinker-7B | 47.68 _{0.14} | 68.67 _{0.06} | 58.44 _{0.16} | 60.07 _{0.15} | 83.77 _{0.04} | <u>81.15</u> _{0.00} | 62.52 |
| R1-Onevision-7B | 31.54 _{0.10} | 49.87 _{0.46} | 51.50 _{0.83} | 50.19 _{0.18} | 72.50 _{0.32} | 54.19 _{0.79} | 46.88 |
| ViGoRL-7B-Spatial | 36.97 _{0.21} | 65.67 _{0.55} | 58.44 _{0.68} | 58.65 _{0.49} | 82.08 _{0.07} | 77.49 _{0.74} | 60.68 |
| VL-Rethinker-7B | 39.84 _{0.26} | 68.50 _{0.39} | 65.00 _{0.98} | <u>61.57</u> _{0.15} | <u>84.40</u> _{0.08} | 64.57 _{1.08} | 60.99 |
| Vision-G1 | 46.88 _{0.21} | 69.76 _{0.16} | 62.67 _{0.00} | 64.93 _{0.00} | 86.55 _{0.08} | 79.93 _{0.25} | 63.26 |
| Vision-R1 | 39.75 _{0.16} | 67.41 _{0.06} | 58.45 _{0.32} | 60.51 _{0.08} | 79.95 _{0.18} | 78.18 _{0.25} | 58.86 |
| TreeVGR-7B | <u>47.29</u> _{0.38} | 67.58 _{0.11} | 62.11 _{0.79} | 60.26 _{0.16} | 74.80 _{0.07} | 83.60 _{0.49} | 61.11 |
| ThinkLite-7B | 45.36 _{0.43} | 69.37 _{0.06} | 66.44 _{0.42} | 62.19 _{0.32} | 86.85 _{0.04} | 80.28 _{0.25} | 62.61 |

Table 2: Accuracy of SOTA MRMs on 13 spatial benchmarks. The top two rows shows performance of the base model Qwen2.5-VL-7B, which is competitive with MRMs trained to perform multimodal reasoning. We identify that open-source MRMs do not exhibit generalized spatial intelligence beyond their base model.

| Models | Random | Qwen2.5 _{cot} | Qwen2.5 | GThinker | R1-Ov | ViGoRL | VL-Re. | Vision-G1 | Vision-R1 | TreeVGR | ThinkLite |
|------------|--------|------------------------|--------------|----------|-------|--------|--------|-----------|-----------|---------|-----------|
| No-Image | 38.83 | 37.45 | 38.59 | 44.17 | 28.1 | 43.18 | 41.26 | 44.46 | 41.15 | 41.91 | 42.48 |
| No-Image++ | - | 43.4 | 76.41 | 5.55 | 11.22 | 30.95 | 47.73 | 25.28 | 7.29 | 11.35 | 36.00 |

Table 3: Results of two variant of the No-Image ablation, where the images are replaced with an uninformative full gray image. For No-Image, MRMs show much higher average performance, indicating their ability to shortcut an answer just from the question (random is better). For No-Image++ (higher better), where a “cannot determine” option is added, we find Qwen_cot as well as MRMs still choose other options as they are biased by the text trace.

models produce significantly shorter traces (~ 350 characters for GPT-5-mini vs. ~ 3600 characters for Qwen3-VL-8B-Thinking), and (ii) proprietary traces lack the reflective phrases (e.g., “wait”, “let me reconsider”) and repetitive looping commonly observed in open-source MRMs. An example output is shown in Appx. § C.3. We hypothesize that this conciseness helps proprietary models avoid the hallucination-inducing verbosity that harms open-source models, though their training details remain opaque. These observations suggest that the quality and conciseness of reasoning traces, rather than their mere presence, may be key to preserving spatial reasoning performance under CoT prompting.

(ii) RL-trained Multimodal Reasoning Models Underperform their Backbone. We present the accuracies of eight open-source MRMs on 13

datasets in Tab. 2. The first two rows indicate backbone results, where *cot* indicates evaluation under CoT prompt. We observe that the non-CoT Qwen2.5-VL backbone shows strong average performance of 62.68%. Surprisingly, despite extensive SFT and RL training designed to enhance visual reasoning, seven out of the eight MRMs fail to surpass this baseline. Even models explicitly finetuned for spatial tasks ViGoRL (-2%) & TreeVGR (-1.57%) underperform the backbone. Vision-G1 is the only exception, and outperforms the backbone by $+0.6\%$. However, as we show in the paragraph below, Vision-G1 also posts the highest No-Image accuracy among all MRMs (Tab. 3), suggesting its performance may stem partly from exploiting dataset shortcuts and question-only textual cues rather than grounded visual reasoning.

(iii) Reasoning Models Show Over-reliance on Text Rationale. We highlight a crucial shortcoming of text-only CoT reasoning, where we observe an over-reliance on the text modality which leads to hallucination of visual content. We perform a *No-Image* ablation, where we pass an uninformative fully gray image (of the same size and aspect ratio as the original image) as input along with the question. The first row of Table 3 reports the average accuracy across all 13 datasets. We observe that MRMs perform significantly better than random guessing (e.g., GThinker achieves 44.17%), indicating they can "shortcut" the answer by ignoring visual content and relying solely on question text, options, and world-knowledge priors.

To verify this behavior stems from hallucination rather than lucky guessing, we introduce the *No-Image++* setting, which retains the gray image but appends a "Cannot determine from the image" option as the ground truth. As the second row of Tab. 3 shows, the Non-CoT backbone correctly abstains on 76.41% of samples, whereas MRMs fabricate reasoning traces supporting one of the original options. Fig. 2 provides a qualitative example, and Fig. 3 shows the effect generalizes across three MLM families.

ViGoRL hallucination under No-Image++ (blank image)

Question (CVBench2D): "Where is the cave located with respect to the trees? Options: A. above, B. below, C. Cannot determine from the image"

Output: <think> The cave entrance appears to be at the bottom of the image, while the trees are higher up, suggesting that the cave entrance is below the trees. (300, 450) ...The cave entrance is clearly situated at the bottom of the image, which is below the trees. (250, 550) </think> <answer> below </answer>

Note: The MRM hallucinates coordinates and spatial relationships using world knowledge.

Figure 2: Qualitative examples of No-Image++ failure mode. ViGoRL hallucinates detailed spatial reasoning for a blank image in the No-Image++ setting.

CoT-induced hallucination is not Qwen-specific.

To rule out a backbone-specific artifact, we repeat the No-Image++ probe on two further backbones from different MLM families: InternVL3-8B and LLaVA-v1.6-Mistral-7B. As Fig. 3 shows, all three Non-CoT backbones correctly abstain on most samples (avg. 76.41% Qwen2.5-VL, 85.52% LLaVA-v1.6, 68.07% InternVL3), but switching to CoT prompting reduces abstention in every family (by -33.0%, -52.8%, -17.5% respectively), confirming that the failure arises from the CoT paradigm itself.

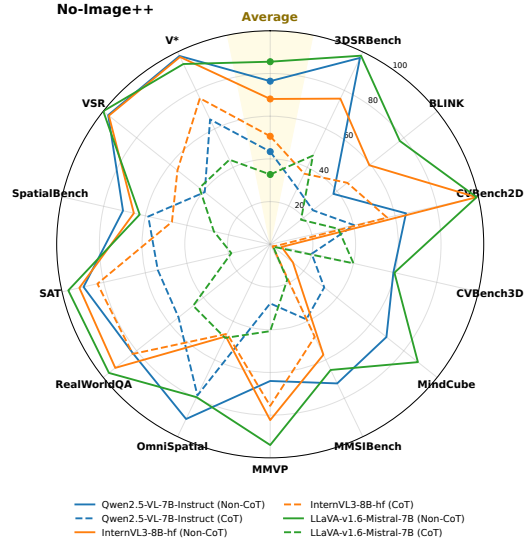


Figure 3: No-Image++ accuracy across 13 benchmarks for three MLM backbones under Non-CoT (solid) and CoT (dashed) prompting. CoT causes all three families (Qwen2.5-VL, InternVL3, LLaVA-v1.6) to hallucinate instead of selecting the "Cannot determine" option.

| Model | Avg. CoT | Avg. Non-CoT |
|---------------------------------|---------------|---------------|
| GPT-4o (Hurst et al., 2024) | 65.55 (+0.50) | 65.05 |
| GPT-4.1-mini (OpenAI, 2025) | 67.79 (+0.39) | 67.40 |
| GPT-5 (Singh et al., 2025) | 69.00 | 69.65 (+0.65) |
| GPT-5-mini (Singh et al., 2025) | 69.86 (+0.08) | 69.78 |
| GPT-5-nano (Singh et al., 2025) | 60.63 | 61.86 (+1.23) |

Table 4: CoT vs Non-CoT proprietary models. The dataset-wise results are shown in Appx. Table 9.

4 Conclusion

We show that the success of reasoning models in logic and math does not yet extend to the spatial domain. Our benchmarking of twenty-one models across thirteen datasets reveals that CoT prompting consistently degrades spatial reasoning performance, with specialized MRMs frequently underperforming their own base models. Crucially, our *No-Image++* analysis identifies the mechanism behind this failure: current reasoning chains tend to hallucinate visual information based on textual priors rather than engaging in grounded perception. Our analysis of proprietary models further suggests that concise, non-repetitive reasoning traces may mitigate this degradation. Our work highlights the need for vision-centric training paradigms for MRMs. Promising future directions include (i) test-time visual verifiers that evaluate each reasoning step against image evidence and trigger backtracking on incorrect visual claims, and (ii) visual process reward models that incentivize grounded, perception-first reasoning during training.

Limitations

In this work, we have sought to cover a broad range of visual spatial reasoning datasets and R1-style MRMs. However, we do not claim that the 13 datasets included represent the entirety of the visual spatial reasoning domain. Given the current landscape of MRMs, it is challenging to completely isolate all confounding factors that may lead to performance improvements or declines across these datasets. We note that proprietary model training details remain opaque, limiting deeper analysis of their behavior. We believe this study offers a solid foundation for future research to further explore vision-centric reasoning paradigms.

Acknowledgments

Sai Srinivas Kancheti would like to thank Microsoft Research India for the MSR India PhD Award and grant, under which this work was carried out. We thank the anonymous reviewers for their valuable feedback that improved the presentation of this paper.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, and Wenbin Ge et al. 2025a. [Qwen3-vl technical report](#). *ArXiv*, abs/2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *ArXiv*, abs/2502.13923.
- Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xieoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2025. [Spatialbot: Precise spatial understanding with vision language models](#). In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37 – 46.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, and Jiaqi et al. Wang. 2024. [Vlmevalkit: An open-source toolkit for evaluating large multi-modality models](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. [Blink: Multi-modal large language models can see but not perceive](#). In *Computer Vision – ECCV 2024*, pages 148–166, Cham. Springer Nature Switzerland.
- Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, and Ruoyu Zhang et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *ArXiv*, abs/2501.12948.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaoshen Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. [Vision-r1: Incentivizing reasoning capability in multimodal large language models](#). *ArXiv*, abs/2503.06749.
- OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and Aleksander Mkadry et al. 2024. [Gpt-4o system card](#). *ArXiv*, abs/2410.21276.
- Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. [Omnispacial: Towards comprehensive spatial reasoning benchmark for vision language models](#). *arXiv preprint arXiv:2506.03135*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Nathan Lambert, Jacob Daniel Morrison, Valentina Pyatkin, Shengyi Huang, and Hamish Ivison et al. 2024. [Tulu 3: Pushing frontiers in open language model post-training](#). *ArXiv*, abs/2411.15124.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. [Llava-onevision: Easy visual task transfer](#). *ArXiv*, abs/2408.03326.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. [Visual spatial reasoning](#). *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 2024. [3dsr-bench: A comprehensive 3d spatial reasoning benchmark](#). *arXiv preprint arXiv:2412.07825*.
- OpenAI. 2025. [Introducing GPT-4.1 in the API](#). <https://openai.com/index/gpt-4-1/>. Accessed: 2026-02-20.
- Arijit Ray, Jiafei Duan, Ellis L Brown II, Reuben Tan, Dina Bashkurova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. 2025. [SAT](#):

- Dynamic spatial aptitude training for multimodal language models. In *Second Conference on Language Modeling*.
- Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. 2025. Grounded reinforcement learning for visual reasoning. *ArXiv*, abs/2505.23678.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.
- Aaditya K. Singh, Adam Fry, Adam Perelman, Adam Tart, Adithya Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, and Akhila Ananthram et al. 2025. OpenAI GPT-5 system card. *ArXiv*, abs/2601.03267.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356. Curran Associates, Inc.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.
- Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijiang Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. 2025a. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *ArXiv*, abs/2507.07999.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025b. V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, and Zhe Chen et al. 2025c. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv*, abs/2508.18265.
- Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Qinghong Lin, Furong Huang, and Lijuan Wang. 2025d. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *ArXiv*, abs/2504.07934.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.
- Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094.
- xAI. 2025. Grok-1.5 vision. <https://huggingface.co/datasets/xai-org/RealworldQA>. License = CC BY-ND 4.0.
- Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. 2025a. Mmsi-bench: A benchmark for multi-image spatial intelligence. *ArXiv*, abs/2505.23764.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyang Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615.
- Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. 2025. Spatial mental modeling from limited views. *Preprint*, arXiv:2506.21458.
- Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2025. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *ArXiv*, abs/2508.12680.
- Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, Ming Tang, Minghui Qiu, and Jinqiao Wang. 2025. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. *Preprint*, arXiv:2506.01078.

A Prompts

In this section we present the dataset prompts as well as different system prompts used in the baselines.

A.1 System Prompts

Base prompt. This is the simple no-thinking prompt used by Qwen2.5-VL-7B. You are a spatial-reasoning assistant. The user asks a question, and the Assistant solves it.

CoT prompts. We give the list of CoT system prompts we use to evaluate the MRM baselines. Below we present prompts used by GThinker, R1-Onevision, ViGoRL-Spatial, VL-Rethinker, Vision-G1, and Vision-R1 respectively.

GThinker: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think>` and `<answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>``<answer>` answer here `</answer>`. In the reasoning process enclosed within `<think>` `</think>`, each specific visual cue is enclosed within `<vcues_*>...</vcues_*>`, where `*` indicates the index of the specific cue. Before concluding the final answer, pause for a quick consistency check: verify whether the visual cues support the reasoning and whether each step logically follows from what is seen. If correct, conclude the answer; otherwise, revise the visual cues and reasoning, then conclude.

R1-Onevision: You are a spatial-reasoning assistant. The user asks a question, and the Assistant solves it. First output the thinking process in `<think>` `</think>` tags and then output the final answer in `<answer>` `</answer>` tags.

ViGoRL-Spatial: A conversation between User and Assistant. The User asks a question, and the Assistant solves it.

The Assistant systematically reasons through the problem step by step by checking and verifying possible solutions and image regions, while grounding reasoning steps to specific objects and their relationships in the image using (x,y) coordinates. There may be one image or two images concatenated together, in which case the Assistant must compare the spatial relationships between the two images. All reasoning processes must be enclosed within a single set of `'<think>'` tags, and reasoning steps must include specific reference coordinates. For example, `<think>` {Reasoning text}. {Further reasoning text} {more reasoning} `</think>` The final answer should be enclosed in `'<answer>'` tags in the format: `<answer>` {text of selected answer choice} `</answer>` The Assistant must help the user identify the correct answer choice from the options provided. - If the correct answer is unclear, select the most relevant option based on the spatial relationships and dynamics within the image. - The Assistant should verify each step and check multiple possible solutions before selecting the final answer.

VL-Rethinker: Please think step by step, and **regularly perform self-questioning, self-verification, self-correction to check your ongoing reasoning**, using connectives such as "Wait a moment", "Wait, does it seem right?", etc. Remember to put your final answer within `□`.

Vision-G1: You **FIRST** think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process **MUST BE** enclosed within `<think>``</think>` tags. The final answer **MUST BE** put in `□`.

Vision-R1: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process

and answer are enclosed within `<think>` `</think>` and `<answer>` `</answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>``<answer>` answer here `</answer>`.

A.2 Dataset Prompts

We use prompts from the respective papers for OmniSpatial, MindCube and Spatial457 as recommended.

OmniSpatial: Task \n -- \n You will receive 1. **Image** - a single RGB frame depicting a scene. \n 2. **Question** - a natural-language query about spatial relationships between objects in the image. \n 3. **Options** - ≥ 2 answer candidates, each tagged by a capital letter (A, B, C, D...).\n Based on the image and question, provide your answer. Always ground your answer in the visual evidence; do not hallucinate unseen objects. If uncertain, pick the most plausible option—never refuse or reply “insufficient information.”

MindCube: Your task is to analyze the spatial arrangement of objects in the scene by examining the provided images, which show the scene from different viewpoints.

A.3 LLM Judge Scoring Prompts

We show the prompts we use for judging MLM generations.

MCQ Scoring: You are a helpful assistant. \n\n Your task: given (1) a free-form “Response” and (2) a list of “Options”, decide which option the response most likely corresponds to and return the option letter. If no option clearly matches, output “0”. \n\n Inputs: - Response: free-form text that may include a letter, a phrase, or an explanation. - Options: A series of choices, each starting with a single uppercase letter followed by “.”, one option in each line.\n\n Output format: - STRICTLY OUTPUT EXACTLY ONE CHARACTER: a single uppercase option letter from the

allowed set, or “0”.\n - Do not output any explanation, spaces, punctuation, or additional text.\n\n Rules:\n 1) If the response explicitly names exactly one letter (patterns like “A”, “A””, “Option A”, “Answer is C”), return that letter immediately.\n 2) Only evaluate the explicitly provided choice. If the response is long and complex without an explicit final choice, return “0”.\n 3) If multiple choices appear in the response, the last unambiguous one is the final choice.\n 4) Never judge factual correctness—only map the response to the best matching option letter from the given options.\n 5) If no explicit letter can be extracted from the response, compare the response’s meaning to option texts. If exactly one option clearly restates or is a synonym/number/name/unit match for the response, return its letter. (Example: response “1956” matches option “B. 1956”)\n 6) If the response uses standard MCQ phrases such as “none of the above” or “all of the above” and a matching option exists, map them. If there is no matching option, output “0”.\n 7) If the response contains both an explicit letter and a conflicting phrase, prefer the explicit letter. If conflicts remain or are unclear, output “0”.\n 8) If the response says “I don’t know”, “Cannot determine”, or similar, output “0”.\n\n - Example 1\n Response:\n Rome\n Options:\n A. Paris\n B. Berlin\n C. Rome\n D. Madrid\n\n Output -> C\n\n - Example 2\n Response:\n I don’t know\n\n Options:\n A. Glucose\n B. Fructose\n C. Sucrose\n D. Lactose\n\n Output -> 0\n\n - Example 3\n Response:\n A. B\n\n Options:\n A. B\n B. D\n C. A\n D. C\n\n Output -> A\n\n

VQA Scoring: You are a helpful assistant.\n\n Task: Given a short free-form “Response” and a gold-standard “Gold”, decide if the Response expresses the SAME answer as Gold. Output “1” for match, “0” otherwise.\n\n Inputs:\n - Gold: the gold-standard answer which is either (i) a short phrase, (ii) an integer, or (iii) “Yes”/“No”.\n -

Response: a few words or a short phrase, possibly will include reasoning steps before the final answer.
 Output format:
 - STRICTLY OUTPUT EXACTLY ONE CHARACTER: "1" if matching, "0" if not.
 - Do not output any explanation, spaces, punctuation, or additional text.
 Rules:
 1) Compare only the final answer in the Response to Gold. Ignore any reasoning steps or intermediate answers present in the Response.
 2) If multiple conflicting answers or uncertainty like "I don't know" appear in the Response, output "0".
 3) Do not use external knowledge; judge only based on the text in Gold and Response.
 4) Punctuation, grammar, and minor spelling errors should be ignored. uppercase/lowercase differences should be ignored. hyphen and underscore are ignored. For ex, "double-bus" and "double bus" are considered the same. synonyms of "Yes"/"No" like "Y"/"N", "True"/"False" must be considered the same. word representations of numbers like "one"/"two"/"three" must be considered the same as "1"/"2"/"3".
 5) Core concept and critical attributes must match. For example, "New York City" and "New York State" do not match. Other examples of non-matches are "bus" vs "double bus"; "red" vs "light red"; "dog" vs "golden retriever"; "apple" vs "green apple".
 6) If the response says "I don't know", "Cannot determine", or similar, output "0".
 Examples:
 - Gold: Double Bus | Response: This is a bus -> 0
 - Gold: Double Bus | Response: I can see a double-bus -> 1
 - Gold: Yes | Response: Y -> 1
 - Gold: 10 | Response: ten -> 1
 - Gold: red | Response: light red -> 0
 - Gold: stop sign | Response: a stop sign on a pole -> 1
 - Gold: person | Response: man -> 0
 Now read the following Gold and Response and output exactly one character: "1" or "0".

B Expanded Tables

Below we present the full table for the No-Img Ablation containing results for all 14 datasets.

In the expanded table below, we provide dataset-wise number for CoT vs Non-CoT performance of MLMs of various backbones and sizes.

Gthinker-7B: visual clues (description of salient part in text) within `<vcues_*></vcues_*>` tags, encourage rethinking (to enable reflection & relook). Build 7k sample CoT dataset using 'pattern-guided cold start' (from ScienceQA, M3CoT, Math, Sherlock etc). CoT data generated using a cascade of MLMs ensuring some samples have rethinking stages. RL data sources are very diverse and general (from llava-o1, r1onevision, mm-eureka). 4K samples are pick from RL sources post clustering to enforce diversity. RL & SFT data sources are different.

ViGoRL-7B: base models do not perform visual verification & don't perform backtracking/reflection. Vanilla GRPO also does not incentivize this behavior. Two step process, i) warm-start CoT SFT: MCTS to generate grounded reasoning steps, where each reasoning step anchors though to image coordinates $\langle s_t, (x_t, y_t) \rangle$. Pref MCTS over linear rollouts to enforce exploration and corrective reflection. Qwen2.5-VL-72B teacher is used to generate about 20k reasoning traces from 1400 images of SAT data (from a total of 32k images) ii) spatially grounded RL: Entire training set of 32k SAT questions.

Vision-G1: multi-domain data curation. Training data spanning many domains (all domains of VisCoT) is collected. The training sources have cross image reasoning data IconQA, NLVR2, ImageCode and spatial reasoning datasets VQA-AS, Super-CLEVR. Multi-round RL with data curriculum is used (i.e after every round of RL data selection is performed to discard low quality data). Influence function based selection is done using LESS (LESS: Selecting Influential Data for Targeted Instruction Tuning). For difficulty based filtering, use prev round checkpoint to generate k rollouts for each sample, and retain those with avg. acc between 0.2&0.8 (dicard too easy and too hard samples). First IF selection gives 40K training samples, then prev round checkpt is used to perform difficulty-based filtering, on which current round RL training is done. Training done for 3 rounds. Unclear if selection & filtering is done fresh over entire data for every round, or selection done once

| Models | 3DSRBench | BLINK | CV-Bench | | MindCube | MMSIBench | MMVP |
|------------------------------|-----------|-------|----------|-------|----------|-----------|-------|
| | | | 2D | 3D | | | |
| Qwen2.5-VL-7B _{cot} | 49.37 | 38.03 | 39.08 | 58.42 | 28.00 | 25.50 | 49.00 |
| Qwen2.5-VL-7B | 51.10 | 38.51 | 29.14 | 55.92 | 32.10 | 25.70 | 50.00 |
| GThinker-7B | 49.72 | 38.66 | 54.87 | 60.50 | 40.95 | 24.20 | 49.67 |
| R1-Onevision-7B | 38.20 | 26.67 | 30.53 | 46.58 | 25.81 | 12.60 | 23.67 |
| ViGoRL-7B-Spatial | 47.70 | 39.77 | 55.29 | 65.00 | 39.90 | 26.30 | 48.33 |
| VL-Rethinker-7B | 51.75 | 40.08 | 36.23 | 59.42 | 34.38 | 24.40 | 49.67 |
| Vision-G1 | 51.85 | 38.66 | 43.60 | 60.17 | 42.57 | 28.90 | 50.00 |
| Vision-R1-7B | 48.23 | 34.88 | 30.18 | 57.92 | 35.24 | 21.70 | 48.00 |
| TreeVGR-7B | 48.61 | 38.93 | 34.49 | 56.25 | 42.48 | 25.00 | 47.33 |
| ThinkLite-VL-7B | 51.39 | 40.24 | 34.77 | 60.08 | 32.10 | 26.00 | 50.00 |

| Models | OmniSpatial | RealWorldQA | SAT | SpatialBench | VSR | V*Bench | Avg. |
|------------------------------|-------------|-------------|-------|--------------|-------|---------|-------|
| Qwen2.5-VL-7B _{cot} | 14.61 | 36.99 | 44.00 | 31.53 | 49.75 | 22.51 | 37.45 |
| Qwen2.5-VL-7B | 15.26 | 36.73 | 47.67 | 35.63 | 48.85 | 35.08 | 38.59 |
| GThinker-7B | 37.70 | 44.18 | 46.67 | 38.62 | 50.25 | 38.22 | 44.17 |
| R1-Onevision-7B | 20.81 | 28.37 | 41.00 | 31.53 | 11.78 | 27.75 | 28.10 |
| ViGoRL-7B-Spatial | 24.07 | 40.65 | 49.00 | 36.94 | 48.04 | 40.31 | 43.18 |
| VL-Rethinker-7B | 21.98 | 42.09 | 51.33 | 37.69 | 50.16 | 37.17 | 41.26 |
| Vision-G1 | 37.90 | 44.05 | 51.00 | 42.72 | 51.47 | 35.08 | 44.46 |
| Vision-R1-7B | 31.64 | 43.27 | 55.67 | 39.37 | 50.08 | 38.74 | 41.15 |
| TreeVGR-7B | 39.92 | 41.31 | 51.67 | 36.01 | 50.41 | 32.46 | 41.91 |
| ThinkLite-VL-7B | 32.22 | 43.14 | 55.33 | 38.25 | 51.55 | 37.17 | 42.48 |

Table 6: Dataset wise expanded results for the No-Image ablation

GPT-5-mini (concise, no reflection; ≈350 chars).

Output: <think> Look at the scene: the catcher is crouched at home plate holding the baseball glove low near the ground. The person in red with the red hat is standing behind near the dugout/stands, clearly higher above the field. So the red hat is at a higher 3D location than the glove. </think>
<answer> B. red hat of the person in red in the back </answer>

Qwen3-VL-8B-Thinking (long, reflective, option-matching; ≈3600 chars).

Output: <think> So, let’s figure out which object is higher in 3D location. First, let’s look at the image. The baseball glove is held by the catcher, who is crouched behind home plate. The red hat of the person in red in the back—those are the people in the dugout or stands, right? The catcher is on the field, closer to the camera, but the red hats are on the players in the background. Wait, in terms of vertical position (height), the red hats are on the heads of people standing, so they might be higher than the catcher’s glove. Wait, the catcher is crouched, so his glove is lower. The people in the back (like the ones in red jackets) are standing, so their hats are higher up. Let’s check the options. Option A is baseball glove (catcher’s glove), which is near the ground, and option B is the red hat of

the person in red in the back. Since the people in the back are standing, their hats are higher than the catcher’s glove, which is at a lower height. So B is higher. </think>

To determine which object has a higher 3D location, we analyze the vertical positions of the options:

- **Option A (baseball glove)**: The baseball glove is held by the catcher, who is crouched behind home plate. This position places the glove relatively low in the scene, close to the ground.

- **Option B (red hat of the person in red in the back)**: The “person in red in the back” refers to individuals in the dugout or stands. These individuals are standing upright, so their hats are positioned higher in the vertical plane compared to the crouched catcher’s glove.

Since the red hat is on a standing person (higher vertical position) and the baseball glove is on a crouched catcher (lower vertical position), **Option B** has a higher 3D location.

<answer> B. red hat of the person in red in the back </answer>

Observations. The GPT-5-mini trace is a single direct chain of perception-to-conclusion with no reflective phrases. The Qwen3-VL-Thinking trace is an order of magnitude longer, contains two explicit “Wait”-triggered re-examinations, redundantly enu-

| Models | 3DSRBench | BLINK | CV-Bench | | MindCube | MMSIBench | MMVP |
|------------------------------|-----------|-------|----------|-------|----------|-----------|-------|
| | | | 2D | 3D | | | |
| Qwen2.5-VL-7B _{cot} | 29.22 | 25.57 | 40.33 | 19.92 | 32.38 | 39.00 | 27.67 |
| Qwen2.5-VL-7B | 96.90 | 37.87 | 65.23 | 59.17 | 69.62 | 72.30 | 64.00 |
| LLaVA-1.6-7B _{cot} | 46.11 | 18.46 | 32.68 | 40.00 | 1.81 | 18.40 | 40.67 |
| LLaVA-1.6-7B | 98.04 | 77.64 | 99.17 | 59.67 | 88.48 | 65.30 | 94.00 |
| InternVL3-8B _{cot} | 36.71 | 46.19 | 56.19 | 3.08 | 1.43 | 48.10 | 75.67 |
| InternVL3-8B | 75.72 | 59.39 | 99.10 | 5.50 | 13.62 | 57.40 | 82.33 |
| GThinker-7B | 0.25 | 0.16 | 1.04 | 0.17 | 0.76 | 2.60 | 2.67 |
| R1-Onevision-7B | 4.60 | 11.84 | 10.36 | 3.08 | 6.00 | 8.90 | 5.33 |
| ViGoRL-7B-Spatial | 36.01 | 14.41 | 30.88 | 1.08 | 8.38 | 27.40 | 24.00 |
| VL-Rethinker-7B | 47.04 | 10.63 | 49.03 | 60.42 | 28.19 | 41.20 | 39.67 |
| Vision-G1 | 13.07 | 6.68 | 12.17 | 7.08 | 12.86 | 28.10 | 14.00 |
| Vision-R1-7B | 1.61 | 0.47 | 5.56 | 2.50 | 0.67 | 0.90 | 5.00 |

| Models | OmniSpatial | RealWorldQA | SAT | SpatialBench | VSR | V*Bench | Avg. |
|------------------------------|-------------|-------------|-------|--------------|--------|---------|-------|
| Qwen2.5-VL-7B _{cot} | 78.54 | 54.90 | 54.00 | 58.58 | 39.20 | 64.92 | 43.40 |
| Qwen2.5-VL-7B | 90.80 | 82.09 | 89.67 | 70.71 | 97.05 | 97.91 | 76.41 |
| LLaVA-1.6-7B _{cot} | 48.66 | 46.14 | 18.67 | 27.05 | 42.39 | 43.98 | 32.69 |
| LLaVA-1.6-7B | 79.45 | 96.60 | 97.00 | 62.69 | 100.00 | 93.72 | 85.52 |
| InternVL3-8B _{cot} | 46.51 | 82.35 | 83.00 | 47.20 | 55.56 | 75.92 | 50.61 |
| InternVL3-8B | 47.88 | 92.81 | 91.67 | 65.49 | 96.64 | 97.38 | 68.07 |
| GThinker-7B | 23.42 | 4.84 | 24.67 | 1.49 | 0.08 | 9.95 | 5.55 |
| R1-Onevision-7B | 27.27 | 12.55 | 21.33 | 12.69 | 4.58 | 17.28 | 11.22 |
| ViGoRL-7B-Spatial | 58.51 | 37.12 | 36.33 | 25.93 | 46.24 | 56.02 | 30.95 |
| VL-Rethinker-7B | 93.28 | 50.59 | 47.00 | 41.60 | 47.95 | 63.87 | 47.73 |
| Vision-G1 | 57.60 | 37.91 | 49.00 | 21.27 | 18.09 | 50.79 | 25.28 |
| Vision-R1-7B | 16.50 | 12.94 | 10.33 | 5.22 | 15.79 | 17.28 | 7.29 |

Table 7: Per-dataset No-Image++ accuracy (seed 1) for the three backbones in Figure 3 (upper block) and for six open-source MRMs (lower block). For the backbones, rows with subscript *cot* use the CoT system prompt and other rows use the base (Non-CoT) prompt; MRMs are evaluated under their native CoT prompt. Non-CoT backbones correctly abstain on most benchmarks, while both CoT-prompted backbones and MRMs hallucinate non-abstain answers across families.

merates the options twice (inside <think> and again after), and repeats the core claim four times. Both reach the same answer, but the open-source trace’s verbosity and option-matching are representative of the pattern we observe across benchmarks, and are the kind of behaviour that amplifies hallucination under adversarial conditions such as our No-Image++ setting.

| Models | 3DSRBench | BLINK | CV-Bench | | MindCube | MMSIBench | MMVP |
|-------------------------------------|-------------|--------------|----------|--------------|----------|-----------|-------|
| | | | 2D | 3D | | | |
| Qwen2.5-VL-3B _{cot} | 53.38 | 45.13 | 71.21 | 68.00 | 40.38 | 24.60 | 63.67 |
| Qwen2.5-VL-3B | 52.10 | 47.97 | 70.58 | 74.00 | 43.71 | 25.80 | 64.67 |
| Qwen2.5-VL-7B _{cot} | 57.11 | 53.44 | 75.92 | 76.09 | 30.83 | 27.47 | 72.44 |
| Qwen2.5-VL-7B | 55.38 | 56.04 | 77.17 | 83.78 | 35.11 | 26.87 | 75.78 |
| Qwen2.5-VL-72B _{cot} | 61.66 | 58.50 | 78.86 | 85.42 | 39.90 | 29.50 | 76.33 |
| Qwen2.5-VL-72B | 59.74 | 63.07 | 79.90 | 86.00 | 42.48 | 32.90 | 79.00 |
| InternVL3-8B _{cot} | 47.93 | 37.66 | 44.92 | 50.42 | 41.62 | 23.60 | 49.67 |
| InternVL3-8B | 51.19 | 38.82 | 45.06 | 57.08 | 36.57 | 28.10 | 49.67 |
| InternVL3.5-38B _{cot} | 56.10 | 60.23 | 82.06 | 90.17 | 31.43 | 12.70 | 81.00 |
| InternVL3.5-38B | 59.80 | 64.49 | 81.99 | 87.58 | 47.05 | 30.60 | 81.33 |
| LLaVA-1.6-7B _{cot} | 45.01 | 30.04 | 37.41 | 47.75 | 40.10 | 26.50 | 45.00 |
| LLaVA-1.6-7B | 45.55 | 21.20 | 41.03 | 54.50 | 40.29 | 29.50 | 49.33 |
| LLaVA-OV-72B _{cot} | 60.48 | 54.81 | 79.49 | 79.42 | 37.52 | 30.30 | 81.00 |
| LLaVA-OV-72B | 60.15 | 58.55 | 80.46 | 85.17 | 48.57 | 30.20 | 84.00 |
| Qwen3-VL-8B-Thinking _{cot} | 60.67 | 59.13 | 78.65 | 92.75 | 35.14 | 28.70 | 76.67 |
| Qwen3-VL-8B-Thinking | 59.69 | 66.70 | 79.21 | 92.67 | 35.14 | 30.40 | 79.33 |
| Models | OmniSpatial | RealWorldQA | SAT | SpatialBench | VSR | V*Bench | Avg. |
| Qwen2.5-VL-3B _{cot} | 40.77 | 62.35 | 55.00 | 56.72 | 74.88 | 71.20 | 55.95 |
| Qwen2.5-VL-3B | 45.92 | 65.88 | 59.00 | 56.53 | 79.21 | 75.39 | 58.52 |
| Qwen2.5-VL-7B _{cot} | 40.40 | 63.05 | 59.22 | 61.75 | 81.83 | 76.27 | 59.68 |
| Qwen2.5-VL-7B | 45.23 | 69.02 | 63.11 | 62.87 | 85.38 | 79.06 | 62.68 |
| Qwen2.5-VL-72B _{cot} | 47.68 | 71.76 | 67.33 | 68.66 | 85.35 | 67.54 | 64.50 |
| Qwen2.5-VL-72B | 49.51 | 73.73 | 71.00 | 69.40 | 87.64 | 78.01 | 67.11 |
| InternVL3-8B _{cot} | 36.92 | 44.05 | 47.00 | 44.22 | 50.57 | 30.37 | 42.23 |
| InternVL3-8B | 36.07 | 40.39 | 46.33 | 43.66 | 49.51 | 34.03 | 42.81 |
| InternVL3.5-38B _{cot} | 48.08 | 69.80 | 64.33 | 61.38 | 81.10 | 66.49 | 61.91 |
| InternVL3.5-38B | 48.27 | 76.21 | 64.67 | 68.10 | 83.88 | 69.11 | 66.39 |
| LLaVA-1.6-7B _{cot} | 28.70 | 32.29 | 43.00 | 36.38 | 45.50 | 25.65 | 37.18 |
| LLaVA-1.6-7B | 22.50 | 39.35 | 35.00 | 36.19 | 49.26 | 34.55 | 38.33 |
| LLaVA-OV-72B _{cot} | 43.77 | 69.02 | 61.00 | 66.42 | 78.07 | 68.06 | 62.26 |
| LLaVA-OV-72B | 48.27 | 71.76 | 66.00 | 69.22 | 79.71 | 67.54 | 65.35 |
| Qwen3-VL-8B-Thinking _{cot} | 40.90 | 73.73 | 74.00 | 67.91 | 82.82 | 79.58 | 65.43 |
| Qwen3-VL-8B-Thinking | 45.73 | 70.98 | 70.33 | 62.87 | 84.21 | 81.68 | 66.07 |

Table 8: Dataset wise table for averages shown in Figure 1.

| Models | 3DSRBench | BLINK | CV-Bench | | MindCube | MMSIBench | MMVP |
|-----------------------------|-----------|-------|----------|-------|----------|-----------|-------|
| | | | 2D | 3D | | | |
| GPT-4o _{cot} | 63.20 | 61.60 | 78.23 | 86.42 | 43.52 | 34.10 | 84.33 |
| GPT-4o | 61.80 | 65.23 | 75.17 | 85.42 | 47.24 | 34.20 | 84.33 |
| GPT-4.1-mini _{cot} | 64.63 | 63.44 | 79.49 | 87.58 | 50.29 | 30.50 | 83.33 |
| GPT-4.1-mini | 62.19 | 64.18 | 80.32 | 88.58 | 48.86 | 30.40 | 82.00 |
| GPT-5 _{cot} | 66.34 | 63.49 | 81.36 | 90.67 | 35.71 | 31.00* | 87.67 |
| GPT-5 | 67.87 | 65.02 | 80.88 | 90.50 | 42.76 | 26.10 | 86.00 |
| GPT-5-mini _{cot} | 65.93 | 63.23 | 81.92 | 90.08 | 51.05 | 33.40 | 84.67 |
| GPT-5-mini | 66.53 | 64.49 | 80.60 | 90.75 | 52.29 | 34.50 | 82.67 |
| GPT-5-nano _{cot} | 58.12 | 50.08 | 76.29 | 86.75 | 18.38 | 24.00* | 75.67 |
| GPT-5-nano | 60.50 | 53.71 | 76.84 | 88.42 | 27.71 | 21.60 | 72.67 |

| Models | OmniSpatial | RealWorldQA | SAT | SpatialBench | VSR | V*Bench | Avg. |
|-----------------------------|-------------|-------------|-------|--------------|-------|---------|-------|
| GPT-4o _{cot} | 45.73 | 73.59 | 68.67 | 63.99 | 84.45 | 64.40 | 65.55 |
| GPT-4o | 46.44 | 76.60 | 64.33 | 63.25 | 80.93 | 60.73 | 65.05 |
| GPT-4.1-mini _{cot} | 51.08 | 79.48 | 74.33 | 69.96 | 83.80 | 63.35 | 67.79 |
| GPT-4.1-mini | 53.29 | 78.69 | 69.33 | 66.98 | 80.20 | 71.20 | 67.40 |
| GPT-5 _{cot} | 51.71* | 81.31 | 82.00 | 66.23 | 86.74 | 72.77 | 69.00 |
| GPT-5 | 50.95 | 82.75 | 85.00 | 68.66 | 86.74 | 72.25 | 69.65 |
| GPT-5-mini _{cot} | 52.51 | 79.61 | 83.00 | 67.35 | 81.67 | 73.82 | 69.86 |
| GPT-5-mini | 53.16 | 80.26 | 82.33 | 64.93 | 81.91 | 72.77 | 69.78 |
| GPT-5-nano _{cot} | 47.97* | 71.90 | 64.00 | 65.49 | 82.00 | 67.54 | 60.63 |
| GPT-5-nano | 46.51 | 74.12 | 72.00 | 66.60 | 81.26 | 62.30 | 61.86 |

Table 9: Dataset-wise results for the five proprietary GPT models evaluated in Table 4. Rows with subscript *cot* use the CoT system prompt; other rows use the base (Non-CoT) prompt. Starred cells (*) report accuracy computed over the non-refusal subset only (GPT-5 and GPT-5-nano sometimes decline to answer on OmniSpatial and MMSIBench under CoT prompting).

| method | SFT | RL data | RL Algo | Rewards | Training | Other |
|-------------|--------------------------------|--|-------------|-------------------------|--------------------------------------|--|
| GThinker-7B | CoT-SFT on 7K dset | diversity based 4K samples | DAPO | fmt + acc Hybrid | SFT 3 epoch RL 170 steps | tab. 3&4 ablation rethink drops gen perf |
| ViGoRL-7B | MCTS CoT SFT on train subset | full train set | GRPO | format + acc +coord fmt | SFT 3 ep, RL 500 steps, kl_coef 0.01 | Multi-turn RL for vis. search |
| Vision-G1 | no SFT | IF selection (40K), difficulty filtering | GRPO w/o KL | fmt + acc | 3-round RL 25 ep each round | multi-round RL |
| Vision-R1 | CoT SFT 200K Modality-bridging | 10K no filtering | GRPO | fmt + acc | SFT 2 ep, RL 300 iters PTST | two above show cold-RFT possible |

Table 10: Various methodological aspects of baselines