Spatial Mental Modeling from Limited Views

(/pdf?id=VcFCThxpUY)

Baiqiao Yin (/profile?id=~Baiqiao_Yin1), Qineng Wang (/profile?id=~Qineng_Wang1), Pingyue Zhang (/profile?id=~Pingyue_Zhang1), Jianshu Zhang (/profile?id=~Jianshu_Zhang3), Kangrui Wang (/profile?id=~Kangrui_Wang2), Zihan Wang (/profile?id=~Zihan_Wang23), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Keshigeyan Chandrasegaran (/profile?id=~Keshigeyan_Chandrasegaran1), Han Liu (/profile?id=~Han_Liu4), Ranjay Krishna (/profile?id=~Ranjay_Krishna1), Saining Xie (/profile?id=~Saining_Xie2), Manling Li (/profile?id=~Manling_Li1), Jiajun Wu (/profile?id=~Jiajun_Wu1), Li Fei-Fei (/profile?id=~Li_Fei-Fei1)

25 Apr 2025 (modified: 18 Sept 2025) Submitted to NeurIPS 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=VcFCThxpUY) BibTeX

Keywords: Spatial Mental Model, Limited Views, Partial Observation, Spatial Reasoning, Cognitive Map

Abstract:
People intuitively construct mental models of space beyond what they directly perceive, but can large visual-language models (VLMs) do the same with partial observations like limited views? We identify this significant gap for current VLMs via our new MINDCUBE benchmark with 17,530 questions and 2,919 images, evaluating how well VLMs build robust spatial mental models, representing positions (cognitive mapping), orientations (perspective-taking), and dynamics (mental simulation) for what if movements, to solve spatial reasoning on unseen space that lies beyond immediate perception.

We explore three approaches to approximating spatial mental models in VLMs: (1) View interpolation to visualize mental simulation, which surprisingly offers little benefit, highlighting the challenge of reasoning from limited views; (2) Textual reasoning chains, which effectively guide model thinking when supervised; and (3) Structured representations like cognitive maps, where ground truth maps help little, but training VLMs to generate and reason over their own maps yields substantial gains—even if the maps are imperfect. Training models to reason over these internal maps raises accuracy from 38.3% to 61.7% (+23.5%). Adding reinforcement learning further improves performance to 76.1% (+37.8%).

Our key insight is that no scaffolding of spatial mental models, actively construct-ing and utilizing spatial mental representations with flexible reasoning chains or processes, significantly improves understanding of unobservable space.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Supplementary Material: zip (/attachment?id=VcFCThxpUY&name=supplementary_material)
Reviewer Nomination: Qineng Wang (/profile?id=~Qineng_Wang1)
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Deep learning (e.g., architectures, generative models, optimization for deep networks, foundation models, LLMs)
LLM Usage: Proving theorems, Implementing standard methods, Implementing methods where LLM plays an original, non-standard role
Declaration: I confirm that the above information is accurate.
Submission Number: 4015

Paper Decision

Decision by Program Chairs 17 Sept 2025, 08:40 (modified: 18 Sept 2025, 09:22) Program Chairs, Authors

Revisions (/revisions?id=bf72RnWxXP)

Decision: Reject

Comment:

This paper introduces MINDCUBE, a benchmark for testing VLLMs on spatial reasoning under partial observability. Baselines show current models perform poorly; the authors explore vision and text prompting, SFT, and RL, with a two-stage recipe—first supervising "cognitive-map" generation, then RL—yielding sizable gains (e.g., 38.3%→61.7%). The empirical study is careful, covers diverse tasks and models, and reports detailed results.

However, the central novelty remains limited: the method largely combines well-known tools (prompting/SFT/RL) without a clearly new algorithmic idea, and key design choices (prompt consistency, reward specification, training and inference protocol) need tighter justification and ablations. During rebuttal the authors provided many additional results that clarify aspects of the design, but these do not fully resolve the novelty concern raised by reviewer 89VQ; reviewer dbE7 also calls for reorganization and integration of new analyses; reviewer WJos is more positive and feels the scope is now clearer.

Weighing these, my recommendation is Reject: the benchmark is valuable and the empirical gains are strong, but the contribution as written remains primarily empirical and the crucial clarifications live outside the submission. I encourage a revision that sharpens the methodological contribution and integrates the rebuttal analyses.

Summary of Revisions and Responses to Reviewers

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

09 Aug 2025, 01:53 (modified: 09 Aug 2025, 01:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors

Revisions (/revisions?id=ouJLpjpLcK)

Comment:

Summary

We thank the AC and reviewers for the thoughtful feedback and constructive discussion. Below is a concise summary of our thesis, key results, and how we addressed the core concerns raised in the reviews.

Thesis

Current VLMs struggle to reason about **unseen** space from **limited views**. We introduce **MindCube** (17,530 Qs; 2,919 images) to diagnose this gap and study scaffolds that help models **construct and use** spatial mental models. Training models to *generate and reason over their own internal maps* lifts accuracy from **38.3% → 61.7%**, and adding RL reaches **76.1%**; in contrast, *view interpolation* offers little benefit.

Key results

1. **Map-then-reason wins.** Explicit "map-then-reason" substantially outperforms passive inputs/extra pixels; the central insight is that *how* the model thinks matters more than *how much* it sees.
2. **Benchmark difficulty and scope.** MindCube targets language-grounded spatial reasoning under partial observability (not geometric reconstruction); even strong VLMs perform far from perfect, underscoring difficulty and significance.
3. **Reasoning > raw pixels.** "More pixels ≠ better reasoning": view-interpolation brings little improvement; internal map + reasoning brings large gains and aligns map quality with task performance.
4. **Warm-started RL matters.** RL (GRPO) *after* SFT drives the jump to **76.1%** and improves map similarity; RL from scratch helps little.

Rebuttal summary

1. **Clarification on Novelty and Contribution (dbE7, WJos, 89VQ).** We clarified that our primary contribution is the identification and validation of an effective language-reasoning cognitive process for VLMs.
2. **Generalizes across model families (dbE7).** On InternVL3-2B SFT, **Aug-CGMap-FFR-Out** and **Plain-CGMap-FFR-Out** achieve **71.44%** and **73.56%**, respectively, well above alternative settings, which supports cross-VLM generality.
3. **Strengthened Empirical Rigor.** We conducted over 10 new experiments to address reviewer concerns. This included:
   - Re-running experiments with standardized prompts to correct an initial inconsistency (qKVt).
   - Adding statistical analysis by running key experiments with multiple seeds to report mean and standard deviation (89VQ).
   - Performing extensive ablations on RL reward design (WJos, 89VQ), SFT hyperparameters (89VQ), and alternative representation formats (e.g., scene graphs, YAML) (WJos).
   - Providing a systematic failure-case analysis (89VQ) and probing studies on latent representations (qKVt, 89VQ).

4. **Additional SOTA Models Perform Badly on MindCube (5ZSD).** We add o3 and o4-mini performance and find they both perform much worse than humans.

Official Review of Submission4015 by Reviewer qKVt

Official Review by Reviewer qKVt 04 Jul 2025, 05:01 (modified: 24 Jul 2025, 07:41)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qKVt

Revisions (/revisions?id=usAF0i8TkF)

Summary:
This paper introduces MINDCUBE, a new benchmark for spatial reasoning in Vision-Language Models (VLMs) under limited views, and demonstrates that explicitly training VLMs to construct and reason over internal cognitive maps significantly enhances their ability to perform complex spatial tasks.

The MINDCUBE benchmark is a significant and timely contribution to the field, and the systematic investigations on frozen VLMs as well as SFT/RL tuned VLMs provide interesting analysis and insights towards more spatially intelligent AI.

Strengths And Weaknesses:
Strengths: 1. The introduction of MINDCUBE is a significant contribution, providing a well-structured and challenging benchmark specifically designed to test spatial mental modeling abilities in VLMs. 2. The paper conducts a thorough investigation into various scaffolding approaches, from view interpolation and free-form reasoning to structured cognitive maps. This systematic analysis provides valuable insights. 3. The results demonstrate remarkable improvements, particularly when VLMs are trained to generate and reason over their own internal cognitive maps, coupled with reinforcement learning.

Weaknesses:

1. Some of the reported results, particularly those involving free-form reasoning (FFR) outputs, appear counter-intuitive and may stem from implementation issues or inconsistencies in prompt design. These issues cast doubt on the validity of certain performance comparisons. Specifically: • CGMap-In-FFR-Out is intended to produce free-form reasoning (as indicated in Table 2), yet its prompt lacks an explicit [object Object] instruction. In contrast, higher-performing settings like Aug-CGMap-FFR-Out and Plain-CGMap-FFR-Out include this format (Appendix C.3.7–C.3.9). • Appendix D.1 shows that CGMap-In-FFR-Out only outputs the final answer, with no reasoning trace—making it functionally similar to Aug-CGMap-In, which explains their comparable performance in Table 2. • The FF-Rsn configuration is also described as involving free-form reasoning (FFR), but its prompt does not follow the same response format as other FFR settings, complicating fair comparisons.
2. The paper lacks an ablation study of CGMap-In-FFR-Out under fine-tuning configurations, such as keeping the vision encoder frozen while toggling MLP layers (frozen/unfrozen). According to Table 4, generating cognitive maps is not particularly challenging for the MLLM. Adding these variants would make it easier to disentangle the contributions of spatial alignment and free-form reasoning during SFT.

Quality: 3: good
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:
Definition and Interpretation of "Spatial Representations" • What exactly is meant by "spatial representations" in the context of this paper? Are these representations latent (emerging from the model's internal state) or explicit (i.e., the generated or input cognitive maps)? • If the latter, the term "spatial mental modeling" may be overstated, as it relies on human-crafted structure, which introduces strong inductive biases. If the authors intend to argue for latent spatial representations, additional evidence (e.g., linear probing, attention visualization, or activation analysis) would be needed to support that claim.

Limitations:
yes

Rating: 5: Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate-to-high impact on more than one area of AI, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
N/A

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes

Rebuttal by Authors

Rebuttal

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

31 Jul 2025, 00:04 (modified: 31 Jul 2025, 14:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=M8Snp3QGR4)

Rebuttal:

We sincerely thank Reviewer qKVt for their detailed feedback and for recognizing our work as a "significant contribution" with a "well-structured benchmark," "thorough investigation," and "remarkable improvements."

For weaknesses and questions, we address each point below and have updated the manuscript accordingly.

Response to Weaknesses

1. **Regarding the inconsistency in free-form reasoning (FFR) prompts (`CGMap-In-FFR-Out` and `FF-Rsn`).** We thank the reviewer for their careful reading and for identifying this issue.
   - **You are correct; the prompt for `CGMap-In-FFR-Out` was missing an explicit instruction, and the `FF-Rsn` prompt format was different.** We have standardized all FFR prompts and re-run these experiments to ensure a fair comparison.
   - The corrected results are presented below (%).

Frozen VLM Config.	Overall	Rotation	Among	Around
Raw-QA	*37.81*	34.00	36.00	45.20
VI-1	37.90↑	35.50	37.33	41.20
VI-2	37.81–	35.50	36.50	42.80
Aug-CGMap-In	32.00↓	35.00	30.50	33.20
FF-Rsn	**40.48**↑	32.00	36.00	58.00
Aug-CGMap-FFR-Out	**40.57**↑	21.00	43.00	50.40
Plain-CGMap-FFR-Out	**41.33**↑	25.00	39.67	58.40
CGMap-In-FFR-Out	**41.43**↑	37.00	41.67	44.40

   - **Unified Prompt Format:** All configurations involving FFR now use the following unified instruction to elicit a chain of thought:

> Please do step by step reasoning first, then give your final answer. For example, if you think the correct answer is 'A. Above' from 'A. Above B. Under C. Front D. Behind', your response should be this format: '[object Object](replace with your reasoning here)[object Object][object Object]A. Above[object Object]'.

   - **Key Takeaways Remain Unchanged:** With the corrected experiments, our central conclusions still hold:
     - Explicit reasoning is crucial for improving performance.
     - Cognitive maps can help guide the reasoning process.
     - Passive structures (like maps as input) alone and visual continuity offer little benefit.
2. **Regarding the missing fine-tuning ablation for `CGMap-In-FFR-Out` .** We thank the reviewer for this excellent suggestion to further analyze the model's components.
   - We have conducted the requested ablation in our submission for `Raw-QA` (Table 6 in Appendix E.4).
   - We also ran SFT for the `CGMap-In-FFR-Out` by training all parameters versus freezing the vision encoder. The results are as follows:

SFT Configuration	Overall Accuracy (%)
CGMap-In-FFR-Out (Tune All Parts)	55.90
CGMap-In-FFR-Out (Freeze Vision Encoder)	56.67

   - Both tables confirm that the primary challenge lies in teaching the LLM to effectively reason over the provided structured spatial information.
   - We will add this valuable study to the final paper.

Response to Questions

1. **On the definition of "spatial representations."** We thank the reviewer for this important question, which allows us to clarify a core concept.
   - In this work, "spatial representations" are abstractions of an environment that encode the relative configuration of objects and viewpoints.

- Following recent approaches [1, 2], we leverage *explicit scaffolds* that we ask the model to generate 2D cognitive maps and/or free-form reasoning.
- We also use these scaffolds to help VLMs cultivate an internal, flexible understanding through SFT and RL.

2. **On the justification for "spatial mental modeling."**
   - We agree with the reviewer and build on the cognitive science literature, which defines spatial mental models as internal, schematic, and manipulable representations of an environment [3, 4].
   - Our scaffolds are external tools to approximate an internal process. We use explicit structures as a scaffolding mechanism to teach the VLM to *approximate* mental models about reasoning over unseen space, a process it struggles with on its own.
3. **On providing evidence for latent spatial representations.**
   - Despite our working focusing more on *explicit representations*, we are also curious about model's internal features.
   - Experiment Setup: To test whether model maintains latent spatial representations and encodes a viewpoint-invariant “neural line” that signals same object, we assembled 100 object triplets (front / left / right views) from MindCube. We registered PyTorch forward hooks at 10th LLM layer (llm_hidden_states) from our SFT model. Three complementary analyses were performed.
     1. Pairwise similarity: cosine similarity and Pearson r were computed between activations of the same object across views (positive pairs) and between different objects (negative pairs). Both cosine and Pearson similarities for positive pairs exceed those for negative pairs in layer-10 LM, indicating that the model attains a measurable, layer-wise spatial consistency for object identity across viewpoints.

Metric	Positive pairs	Negative pairs
Cosine	0.9651	0.9160
Pearson *r*	0.2750	0.2046

     2. Stable-dimension search: for each token dimension， we computed the variance across the three views. With only 64 stable dimensions (variance 4.15–6.01), cosine similarity across views averages 0.931 ± 0.068, far above random baselines. Low-variance dimensions act as shared neurons, staying nearly constant across viewpoints. Their high pairwise cosine similarity confirms the model has learned an invariant representation.
     3. Probing: a linear classifier trained on activations from front+left views to predict rotation direction (clockwise vs counter-clockwise 90°) reached 0.85 accuracy on held-out objects (chance 0.50), confirming that the stable dimensions generalize to unseen scenes.

---

[1] *Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces*. Yang et al. CVPR 2025

[2] *Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs.* Yeh et al. arXiv:2504.15280

[3] *Mental models in cognitive science.* Laird et al. Cognitive science, 1980

[4] *Mental models: Towards a cognitive science of language, inference, and consciousness* Laird et al. Harvard University Press, 1983.

Replying to Rebuttal by Authors

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

03 Aug 2025, 12:06 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewer qKVt,

Thank you again for your thoughtful and constructive review, and we would like to kindly follow up on your opinions on our rebuttal.

We worked extensively to address the points you raised, and we are keen to ensure our revisions are satisfactory and would be grateful for your feedback when you have a moment.

Thank you for your time and consideration.

The Authors

Official Review of Submission4015 by Reviewer dbE7

Official Review by Reviewer dbE7 03 Jul 2025, 03:55 (modified: 18 Sept 2025, 11:31)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer dbE7
Revisions (/revisions?id=gIghPmgXmB)

Summary:
Paper tackles the challenge of spatial understanding in VLMs from a few limited views of one or many objects with a predefined set of motion maneuvers. Some of the QA types are “what-if” mental simulations from the given view (such as “what if turning to left”), perspective taking (such as “what if taking the sofa’s perspective”) and complex relation reasoning queries.

They show that existing VLMs struggle at many of these task. Then they study three kinds of data structures that are used as input or output scaffolding – interpolated camera views using Stable Virtual Camera, a BEV location encoding of each object (called cognitive map) and free form reasoning (textual reasoning). It appears that camera views are not useful, prompting to output cognitive maps seem to be useful, but inputting cognitive maps and asking to do free form reasoning is not useful.

Finetuning on raw QA pairs increases accuracy but not by a lot, this is their finetuning baseline. Finetuning with template reasoning chains improves the most over the baseline, while maps offer limited benefit. Finetuning while generating both maps and reasoning chain provides the most advantage. Then they additionally apply GRPO over this model and show that this provides the most advantage, in my opinion this is not surprising as prior work [a, b, c] has shown that RL over a good finetuned model sharpens the distribution further towards end-task performance metrics.

[a] Tuning computer vision models with task rewards, PMLR 2023

[b] Visual-RFT: Visual Reinforcement Fine-Tuning, Arxiv 2025

[c] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, Arxiv 2025

Strengths And Weaknesses:
Interesting dataset and benchmark with good amount of experimental results, and is generally well executed. Very thorough analysis of the Qwen model family, good coverage showing that existing VLMs do not well on this task.

Writing Style: The paper is not written in a style that I appreciate (too many, very cluttered graphics used), making it harder to read. I could not get the gist of the work from the figures, looking at every figure individually gave me an information overload. Tables are also hard to follow as every experiment has been marked using abbreviations, maybe a table with properties of the model (trained with SFT/RL, scaffolding type used etc etc) with ✓ and × would have been easier to read and interpret.

Novelty in terms of method: More importantly, I admit the value of the dataset and benchmark -- it is novel and poses questions I have not seen in prior work. I’m not sure about the value of the presented insights (for example, useful insights to me would have been --generalizability to many VLM families, or anything interesting about the architectural choices these VLMs make) or the novelty of the methods used. Maybe the paper is more suited for the benchmark track?

Other weakness: Scaffolding experiments seem to be an artifact of the model used [a, b]. All three considered structures are valid, it’s hard to know if there is specific value in the input/output scaffolding that generalizes across VLMs. I admit interpolated camera views are likely less useful as the information content per token of an interpolated view is low, but it’s unclear why the suggested combination of cognitive map and free form reasoning works. Rest of the experiments are standard applications of existing algorithms.

[a] Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs, Arxiv 2025

[b] Spurious Rewards: Rethinking Training Signals in RLVR, Arxiv 2025

Quality: 3: good
Clarity: 1: poor
Significance: 2: fair
Originality: 2: fair
Questions:

1. Most experiments were done with Qwen-VL-2.5. Why are these results generalizable to all VLMs? What is the significance of these results?
2. What is the novel method or insight that is a hallmark of this work? Maybe the paper is more suited to the datasets and benchmark track.
3. What is the accuracy measure? Is there a way for a VLM to obtain partial scores in the considered QA task?

Limitations:
Yes.

Rating: 2: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The authors have responded to many of my comments, and have shown improvements over the specific queries I have asked -- changing my opinion about this work. However, the simplified synthetic setup of their dataset, and presentation concerns partially remain.

Addendum: The authors in response to one of my questions (do the insights of the paper generalize beyond Qwen models) presented results for InternVL3 [a]. However, the language component of InternVL3 is initialized from a Qwen model for post-training, which is the core component already trained on reasoning traces [b, c] (and is unique among VLM families) -- weakening their claims as only the vision encoder is pre/post trained by InternVL team. Thus, my initial question remains unanswered.

I became aware of this during the AC-Reviewer discussion when I started looking into InternVL3 independently, and I find it very problematic that the authors did not mention this caveat at all!

[a] InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, Arxiv 2025

[b] Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs, Arxiv 2025

[c] Spurious Rewards: Rethinking Training Signals in RLVR, Arxiv 2025

Rebuttal by Authors

Rebuttal
by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))
31 Jul 2025, 07:10 (modified: 31 Jul 2025, 14:42)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=BUiheVUSvu)

Rebuttal:
We sincerely thank Reviewer dbE7 for their detailed feedback and for finding our dataset and benchmark novel, and our experiments well-executed and thorough.

For weaknesses and questions, we address each point below and have updated the manuscript accordingly.

Response to Weaknesses

1. **On Clarity of Presentation.** We sincerely thank the reviewer for their valuable and constructive feedback. We agree that clarity is paramount.
   - The density of the figures and tables was a difficult trade-off made to present a comprehensive set of results within the strict 9-page limit.
   - We will redesign and simplify key figures to highlight the core concepts for better clarity.
   - We will revise all tables and adopt the reviewer's excellent suggestion by adding a new summary table using checkmarks (✓/×).
2. **On Limited Novelty of Methods and Insights.** We thank the reviewer for acknowledging the novelty and value of our MindCube benchmark. We respectfully disagree regarding the novelty of our methods and insights, and we appreciate the opportunity to clarify our core contributions.
   - Other than the benchmark, another primary contribution is identifying an effective spatial cognitive process (representation) for VLMs, not inventing a new algorithm.
   - The central question of our work is how VLMs can approximate spatial mental models from limited views.
   - Through our extensive experiments, our key insight that the synergistic "map-then-reason" process significantly outperforms other approaches is a foundational contribution to the underexplored field of spatial reasoning in VLMs.
   - The novelty lies in discovering *what* to teach the model to think, rather than *how* to tune it.
   - Our core "map-then-reason" insight generalizes across different VLM families. We have conducted new `SFT` experiments on `InternVL3`, and the results below show that the `Aug-CGMap-FFR-Out` and `Plain-CGMap-FFR-Out` settings outperform others by a large margin.

InterVL3-2B SFT Config.	Overall Accuracy (%)
Raw-QA	54.23
FF-Rsn	56.83

InterVL3-2B SFT Config.	Overall Accuracy (%)
Aug-CGMap-Out	52.98
Plain-CGMap-Out	51.63
Aug-CGMap-FFR-Out	**71.44**
Plain-CGMap-FFR-Out	**73.56**

- Our focus on cognitive representations over architecture is a deliberate and foundational choice.
  - While architectural impact is important, our work establishes that the structure of the reasoning process itself is a critical factor for spatial intelligence.
  - There is existing work [1, 2] demonstrating the importance of representations for VLMs.

3. **On Generalizability of Scaffolding Results.** We thank the reviewer for these important questions.
   - Our "map-then-reason" finding is not model-specific and generalizes across VLM families (InternVL3) as highlighted in the above table. It shows that our synergistic approach is highly effective on this different model.
   - The effectiveness stems from a powerful synergy between representation and reasoning.
     - We use these artifacts following [3, 4] to see whether VLMs can reason over unseen space or not; if not, then how to alleviate this.
     - Our experimental analysis (lines 231-239 in our submission) reveals this combination is uniquely effective.
     - This could be because the cognitive map provides a structured "mental workspace," and the subsequent reasoning task forces the model to create a *functionally effective* map optimized for inference.
     - As shown in our training dynamics (Fig. 4), this leads to both higher task accuracy and better generated cognitive map quality.
     - Fundamentally different from passive scaffolding, ur work demonstrates that teaching the model *how to think* (map-then-reason) is far more effective than just giving it *more to see*.

Response to Questions

1. **On Generalizability and Significance of Results.**
   - As detailed in our responses to Weaknesses #2 and #3, new experiments on InternVL3-2B validate that our core finding generalizes across different VLMs.
   - The broader significance is showing *what* we teach VLMs to reason about is as important as *how* we build them.
   - Our work provides strong evidence that a model's ability to solve complex spatial tasks is fundamentally bottlenecked by its reasoning process, not just its architecture.
2. **On the Core Contribution of the Work.** We thank the reviewer for this crucial question. As detailed in our response to Weakness #2,
   - First, we systematically identify and benchmark a critical VLM failure point: reasoning about unseen space from limited views. Our MindCube benchmark is the tool we built to prove this foundational gap.
   - Second, we discover an effective cognitive process representation for VLMs: the synergistic "map-then-reason" paradigm.
   - Finally, we demonstrate a highly effective training paradigm for spatial intelligence using "map-then-reason": a curriculum of SFT followed by RL is a highly effective method for spatial reasoning.
3. **On the Definition of Accuracy and Partial Scoring.**
   - Beyond QA accuracy, we use detailed graph metrics (validity, structural isomorphism, and overall similarity) to evaluate the intermediate cognitive maps. Precise definitions are in Appendix C.2.2.
   - We use partial scoring in our RL framework. Models receive a partial reward of +1 for generating a structurally valid output (e.g., a well-formed map) and a larger reward of +5 for the correct final answer, as described in Appendix F.1 (lines 788-793).
   - Scoring the reasoning chain normally is not a straightforward problem. Popular partial scoring methods like LLM-as-judge have key limitations. While an interesting direction, using an LLM to score reasoning chains can introduce evaluation biases and suffer from a lack of consistency [5, 6]. We believe our current, more objective metrics will provide a more reliable assessment.

---

[1] *Incorporating Structured Representations into Pretrained Vision & Language Models Using Scene Graphs*. Herzig et al. EMNLP 2023

[2] *Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs*. Ranasinghe et al. CVPR 2024

[3] *Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces*. Yang et al. CVPR 2025

[4] *Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs.* Yeh et al. arXiv:2504.15280

[5] *Humans or LLMs as the Judge? A Study on Judgement Bias.* Chen et al. EMNLP 2024

[6] *Evaluating the Consistency of LLM Evaluators.* Lee et al. COLING 2025

Replying to Rebuttal by Authors

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

03 Aug 2025, 12:07 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewer dbE7,

Thank you again for your thoughtful and constructive review, and we would like to kindly follow up on your opinions on our rebuttal.

We worked extensively to address the points you raised, and we are keen to ensure our revisions are satisfactory and would be grateful for your feedback when you have a moment.

Thank you for your time and consideration.

The Authors

Have answered many of my queries

Official Comment by Reviewer dbE7 05 Aug 2025, 04:11 (modified: 05 Aug 2025, 04:19)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=79oVqnVz70)

Comment:

Authors have answered many of my queries regarding the work. I'm happy to note the scaffolding results and the model family results are interesting. I agree that this specific line of spatial reasoning is still underexplored (and mechanisms like visual search [a] for spatial reasoning). Spatial reasoning as a new area of research is gaining traction as a space, maybe "map-then-reason" is one such useful novel mechanism. I hope the authors can improve their presentation and make the further revisions easier to read without information overload. Authors should also look at/cite related work which tackles unstructured variants of this problem [b, c].

[a] https://vstar-seal.github.io/ (https://vstar-seal.github.io/)

[b] https://arxiv.org/abs/2309.02420 (https://arxiv.org/abs/2309.02420)

[c] https://sites.google.com/view/ultrra-wacv-2025 (https://sites.google.com/view/ultrra-wacv-2025)

Thanks for your follow-up response

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

06 Aug 2025, 23:17 (modified: 08 Aug 2025, 18:56)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer dbE7

Revisions (/revisions?id=vck1Vscopt)

Comment:

Thank you for the thoughtful follow-up. We're glad you find the scaffolding and model-family results interesting, and we agree that the spatial reasoning paradigm "map-then-reason" is a promising and underexplored direction. Our current submission targets this setting and, we believe, clearly states the problem, method, and empirical findings.

To better situate our contribution, we will include the suggested work and clarify how our formulation is complementary. We will also tighten wording and make sure the figures and tables are clear to improve readability, where space permits.

Thank you again for the constructive feedback.

The Authors

Official Review of Submission4015 by Reviewer WJos

Official Review by Reviewer WJos 03 Jul 2025, 01:07 (modified: 18 Sept 2025, 11:31)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Revisions (/revisions?id=7kM0gGnuNU)

Summary:
This paper introduces MINDCUBE, a new benchmark that tests how well vision-language models can build and work with spatial mental models when they only get limited visual information.

The key idea is helping these models with what the authors call "cognitive scaffolding," basically giving them tools like view interpolation (imagining what things look like from different angles), cognitive maps (creating mental layouts of spaces), and free-form reasoning (thinking through spatial problems step by step).

The results show that when models use these scaffolding techniques, they get better at spatial reasoning tasks, and this improvement gets even stronger when you train them with reinforcement learning on top of these scaffolds.

Strengths And Weaknesses:

Strengths

- Clear benchmark contribution: MINDCUBE fills a niche by targeting spatial reasoning under partial observability, which is (somewhat) underrepresented in existing VLM benchmarks. The diversity of tasks (rotation, perspective-taking, what-if scenarios) is a strength.
- Empirical validation: The experimental results convincingly show that spatial scaffolding improves model performance, and the reinforcement learning extension adds a nice practical dimension.
- Well-organized and accessible: The paper is clearly written and presents a logical flow from motivation to implementation and evaluation.

Weaknesses

- Misuse of the term “cognitive map”: The implementation of “cognitive maps” as JSON-encoded 2D layouts of objects and camera viewpoints is a significant departure from the concept as defined in cognitive science. Epstein et al. (2017), among others, characterize cognitive maps as flexible, allocentric, and relational structures that support navigation and inference. The paper’s use of the term to describe static schematics risks conceptual confusion.
- Questionable choice of JSON representation: The use of a verbose JSON format for “cognitive maps” seems more like an engineering artifact than a cognitively or computationally principled design. It is unclear why this format is preferred over more compact or differentiable representations (e.g., vectorized scene graphs) that might integrate more naturally with VLM architectures.
- I have to admit this paper comes across as pretty safe and boring. It doesn't make a compelling case for why we need MINDCUBE when we already have benchmarks that test spatial reasoning, and it's not clear what real-world capabilities models would actually gain from getting better at these somewhat artificial tasks (partially because the related works are so poorly discussed). While the methods they propose (view interpolation, free-form reasoning, and reinforcement learning) are reasonable enough, they're not particularly groundbreaking. Most frustratingly, the whole paper feels like they took the easy route instead of tackling the deeper, more fundamental questions about how spatial reasoning actually works, leaving us with yet another incremental benchmark rather than genuine insights into spatial cognition. At the end of the day, sparse view reconstruction has been there for years...It's obvious that the current visual tokenization fails to help in such tasks.

Epstein, Russell A., Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. "The cognitive map in humans: spatial navigation and beyond." Nature neuroscience 20, no. 11 (2017): 1504-1513.

Quality: 3: good
Clarity: 3: good
Significance: 2: fair
Originality: 2: fair
Questions:
Q1: Why JSON for cognitive maps? JSON seems like an odd choice for spatial reasoning. Have you considered more natural alternatives like simply connecting to multiview diffusion models, or scene graphs, that would integrate better with how models actually process spatial information?

Q2: How does this relate to sparse view reconstruction? Computer vision has been doing sparse view reconstruction for years with methods like structure-from-motion. Is MINDCUBE targeting something fundamentally different, or is it essentially these CV tasks with a language layer on top? Why not just use existing 3D reconstruction techniques as scaffolding?

Q3: On the RL reward design: Why use +1 for valid outputs and +5 for correct answers? This seems like a pretty arbitrary choice - how did you settle on these specific values, and did you experiment with other reward structures?

Limitations:

yes
Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
N/A

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The current title and framing, to me, are overstated. The contributions are primarily on the language side, with minimal novelty on the visual front. While the paper touches on an important and challenging problem, it feels like an early step rather than a deep exploration, and I find it difficult to get genuinely excited about the current scope of the work.

> We respectfully doubt that the reviewer has a bias that language-based improvements are less important than visual-based.

I guess authors might argue that I am biased toward vision. I am not. Or they should be responsible for this impression: It's not that I hold a bias against language-based contributions; rather, the way this paper is written gives the strong impression that the authors did not intend to engage meaningfully with the language side of spatial reasoning in the first place.

If the claim is to be contributing to “language-based spatial reasoning”, then why is there absolutely no reference in the related work section? There is an extensive literature on how spatial relations are expressed, inferred, and grounded in language (e.g., via landmarks, trajectors, prepositions, spatial roles). None of this is acknowledged.

To build on the previous discussions, if the paper (1) is intended to (as argued by the authors) and (2) can place greater emphasis on spatial language, with explicit connections (similar to the proposed experiment design I mentioned in the previous thread, but not necessarily follow what I proposed) and proper acknowledgments made, I would be willing to raise my score to borderline accept.

I believe the additional experiment can be completed before the camera-ready after the discussion period, and is much lighter compared to the commitments the authors have made to other reviewers. I acknowledge the value of this benchmark (this work is indeed more suitable for the benchmark track) and the solid efforts made during the discussion phase.

Rebuttal by Authors

Rebuttal
by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))
31 Jul 2025, 07:15 (modified: 31 Jul 2025, 14:42)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=cvDvNZ8A4d)

Rebuttal:
We sincerely thank Reviewer WJos for their detailed feedback and constructive criticism. We are grateful that the reviewer finds our MindCube benchmark contribution clear and fills a niche, our experimental results convincing, and the paper well-organized and accessible.

We believe the primary concerns raised stem from a main **misunderstanding of our work's focus on VLM-centric reasoning** rather than traditional CV tasks. We have addressed all weaknesses and questions below and will update our manuscript to clarify these crucial distinctions.

Response to Weaknesses

1. **On the use of the term "cognitive map".** We agree that precise terminology is crucial and would like to clarify our conceptual grounding.
   - Our use of "cognitive map" follows its operationalization in recent VLM literature [1, 2, 3]. The term has been adopted by recent prominent works to describe explicit, structured spatial representations that enable VLMs to reason about space.
   - Our JSON format has been widely adopted as a computational model of a flexible structure for VLMs.
     - Ours is designed as a bird's-eye view representation that encodes the *relative positions and orientations* of objects [1].
     - Such a representation aligns with functional principles of cognitive maps in cognitive science [4, 5, 6] in high level.
   - Our goal is to equip VLMs with a scaffold that approximates the *functional role of a cognitive map* that enabling explicit reasoning, rather than replicating its exact neurological basis.

2. **On the use of JSON representation. We chose JSON as a principled way to interface with text-native VLMs, a standard practice for eliciting structured outputs.**
   - VLMs fundamentally operate on sequences of language tokens. JSON is a text-based format that fits naturally with the VLMs.
   - JSON provides a structured, computationally effective format for evaluating complex spatial outputs. This text-based approach is one of the standard methods for eliciting structured knowledge from LLMs and VLMs [1, 7, 8 ,9, 10].
   - Differentiable vectorized representations are a promising research direction but are not able to be understood by VLMs. Integrating such formats has been widely recognized as not-working methods yet for current VLMs [1]. The most related work is Spatial Models which show low performance in Table 1.
3. **On the overall contribution and framing.** We believe there is a serious misunderstanding on the topic of the paper. The reviewer claims that the research question has been researched for a long time, but we cannot identify any prior work on such VLM reasoning setting. Most VLM reasoning are still about non-geometric questions. To the best of our knowledge, MindCube is the first benchmark to systematically target VLM reasoning about unseen space given limited views, especially whether such ability can be approximated in the language reasoning space.
   - MindCube's unique contribution is **evaluating VLM reasoning about unseen space**, which is a fundamental capability for any agent operating in a partially observable world.
   - MindCube is very challenging, even SOTA VLMs **perform near-randomly**.
   - The **real-world impact** lies in identifying a trainable reasoning process for VLMs.
     - One core discovery is the effectiveness of the "map-then-reason" trainable paradigm, working for any embodied agent (e.g., in robotics or AR) that must navigate and act within partially observed environments.
     - We test a set on the unseen real-world dataset CoSpace [11] from Google Maps, to evaluate navigation (north, south, east, west).

Method	Accuracy (%)
Base	34.28
Our SFT	35.92

     - The results show that our SFT model improves accuracy on both unseen real-world dataset and unseen task (navigation), highlighting the generalizability and real-world impact of our model.
   - **Our method investigation provides concrete insights, not just scores.**
     - The way of designing a reasoning reward to force VLMs generate thinking tokens is new, improving to 76.09% from 38.27%, we did not organize the paper through a traditional aggregated method session, but walk through readers about why we design the framework like that.
     - Not only the training, but the first systematic study of how different thinking/reasoning help VLMs approximate spatial mental models.
     - One key finding is that training VLMs to actively *generate their own internal maps* is significantly more effective than providing them with ground-truth maps or denser visual information.
   - **Our benchmark provides the empirical foundation for asking deeper cognitive questions.**
     - MindCube is the critical diagnostic tool, offering the first clues into what constitutes an effective internal spatial representation for VLMs.
   - **Our work distinguishes between sparse-view reconstruction (CV) and sparse-view reasoning (VLM).**
     - We agree that sparse-view reconstruction is a well-studied field.
     - However, our focus is on language-grounded reasoning (e.g., "If you turn left and move forward, will you get closer to the bin?"), a distinct and underexplored challenge for VLMs that requires integrating perception, spatial abstraction, and language understanding.
     - Our finding that view interpolation does not help underscores this distinction (Table 3 in our submission); more pixels do not equal better reasoning.
   - **We will revise our related works section.** We thank the reviewer for this suggestion and will update the paper to more clearly position our contributions against both the VLM literature and classic CV to better highlight our novelty.

Response to Questions

1. **Why JSON for cognitive maps?** As detailed in our response to Weakness #2,
   - Our choice of JSON is principled for VLMs as it aligns with their native text-based modality.
   - It is a standard and effective method for evaluating structured outputs from LLMs and VLMs [1, 7, 8 ,9, 10].
   - Alternatives like multiview diffusion models or differentiable representations are an interesting but orthogonal research direction.
     - Our focus is on classic token-based VLMs to reason over unseen space from limited views with natural language.
     - Multiview diffusion models are interesting and great at either generating 2D or 3D representations, yet in most cases, they are not combined with natural language reasoning.

- Vectorized scene graph has great potential for better spatial performance, but it would require designing specialized modules to bridge continuous vector spaces with the VLM's token-based reasoning process.

2. **How does this relate to sparse view reconstruction?** As detailed in our response to Weakness #3,
   - Our work targets a fundamentally different task: language-grounded spatial reasoning, not geometric reconstruction.
   - Classic 3D reconstruction primarily aims for **geometric fidelity**. Our work, in contrast, targets spatial reasoning tasks like perspective-taking, which demand a semantic understanding that geometric models alone are not inherently designed to provide.
   - Our approach is designed to directly leverage the power of mature, high-performing 2D VLMs. Explicit 3D data formats (e.g., point clouds) are not seamlessly compatible with the architectures of these models, posing significant challenges for direct feature integration.
   - Our finding that view interpolation (a form of reconstruction, table 3 in our submission) fails to improve reasoning underscores this distinction.
3. **Q3: On the RL reward design (+1 for valid, +5 for correct)?**
   - Following Deepseek-R1 [12], we assign a small format score to encourage valid-format generation, and a large answer score to encourage correct answer generation.
   - Our method focuses more on finding an effective cognitive process rather than proposing a new RL algorithm. But we also agree that conducting a reward ablation study will make the results more convincing.
     - We explored three reward configurations for the `FF-Rsn` setting.
     - The results are presented in the table below.

RL Reward Setting (format score, answer score)	Acc (%)
Frozen VLM (baseline)	40.57
Our Setting (+1, +5)	50.57
Answer Only (+0, +5)	50.76
Format Only (+5, +0)	45.05

     - This result confirms that assigning the correct answer with a high value is the most critical reward component, which is consistent with both our setting in submission and recent work on RL for LLMs like [12].

[1] *Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces*. Yang et al. CVPR 2025

[2] *CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs*. Cao et al. arXiv:2412.10439

[3] *SpaceR: Reinforcing MLLMs in Video Spatial Reasoning*. Ouyang et al. arXiv:2504.01805

[4] *Spatial Cognition.* Newcombe et al. MIT Press, 2024

[5] *Cognitive maps in rats and men.* EC Tolman. Psychological Review, 1948

[6] *Cognitive maps, cognitive collages, and spatial mental models.* B Tversky. European conference on spatial information theory, 1993

[7] *Training Language Models to Follow Instructions with Human Feedback.* Ouyang et al. NeurIPS 2022

[8] *GPT-4 Technical Report.* OpenAI. arXiv:2303.08774

[9] *Toolformer: Language Models That Can Use Tools.* Schick et al. NeurIPS 2023

[10] *Visual Instruction Tuning.* Liu et al. NeurIPS 2023

[11] *CoSpace: Benchmarking Continuous Space Perception Ability for Vision-Language Models*. Zhu et al. CVPR 2025.

[12] *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* DeepSeek-AI. arXiv:2501.12908

Replying to Rebuttal by Authors

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

03 Aug 2025, 12:07 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewer WJos,

Thank you again for your thoughtful and constructive review, and we would like to kindly follow up on your opinions on our rebuttal.

We worked extensively to address the points you raised, and we are keen to ensure our revisions are satisfactory and would be grateful for your feedback when you have a moment.

Thank you for your time and consideration.

The Authors

Replying to Rebuttal by Authors

Official Comment by Reviewer WJos

Official Comment by Reviewer WJos 04 Aug 2025, 15:02

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thanks for the additional work!

General response

I like the additional experiment on the RL reward design, and would love to see this experiment in the next iteration.

If the authors believe that "one key finding is that training VLMs to actively generate their own internal maps is significantly more effective than providing them with ground-truth maps or denser visual information," I would suggest that **the current title and framing may come across as somewhat overstated**. The contributions are primarily on the language side, with minimal novelty on the visual front. While the paper touches on an important and challenging problem, it feels like an early step rather than a deep exploration, and I find it difficult to get genuinely excited about the current scope of the work.

Follow up question:

The rebuttal makes a fair point that JSON is a natural fit for LLMs and VLMs since it aligns with their token-based processing and is commonly used for structured outputs. That said, it doesn't fully address why other symbolic or vector-based formats (like scene graphs or more compact notations) couldn't also be adapted into token sequences with potentially better semantic efficiency. It would help if the authors provided examples or evidence showing that JSON performs better in practice, or that alternative formats actually lead to worse outcomes when used with current models. That would make the choice of JSON feel more principled rather than just convenient.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer WJos

Mandatory Acknowledgement by Reviewer WJos 04 Aug 2025, 15:30

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Scene graph results

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

06 Aug 2025, 16:16 (modified: 07 Aug 2025, 07:29)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Revisions (/revisions?id=BEOfGEpPyi)

Comment:

We appreciate that reviewer like RL details and will include them in the revision. Our new experiments on scene graphs are out and hope to address the remaining concerns.

1. **Title: Spatial Mental Modeling**

- We respectfully disagree that current title is overstated. Our title **"Spatial Mental Modeling" clearly scopes the paper to learning an "abstraction" of the space**. "Mental model" has been well recognized in the literature [1,2,3] to learn a language-based abstraction to facilitate reasoning, which exactly matches the scope of our work. In fact, if it is visual-front, previous work call it mental imaginary simulation [4] or mental visualization [5].
- We respectfully doubt that the reviewer has **a bias that language-based improvements are less important than visual-based**.
  - For decades, spatial reasoning has been a problem from both vision and language perspectives. Reasoning has always been an important topic in language, and **language-centric spatial reasoning has been well recognized** as a line of research [6,7,8].
  - Simply offering denser visual info is insufficient for spatial reasoning, as detailed in lines 51-53 and Table 3. The core challenge lies in improving the model's ability to reason.
  - Language and Vision are **simply different perspectives** for advancing spatial reasoning. It is beneficial to **encourage both communities** to contribute to this goal, rather than **diminishing the insights emerging from text domain.**
- There are **many questions to be answered** in language-centric spatial reasoning for VLMs, which is why we **systematically explore** them:
  - "What is different between spatial reasoning vs general reasoning in large reasoning models?"
  - "What granularity of abstraction works?"
  - "RL cannot directly help much. What kind of reward design is needed?"
  - "Is a warm-up/cold-start needed before RL? What info is needed as prior?"
- **Language-centric exploration does NOT mean less challenging than visual-based**:
  - We collected **17,530 questions** with 3 view changes, 4 what-if dynamics, 3 types of relation queries, 2 types of perspective taking, 2 types of visual patterns. SOTA models get close-to-random performance.
  - We **exhaustively explored the design space of language-based abstraction** with 12 design choices. From frozen models to SFT / RL, we did struggle for months with minimal improvements on what signal can make correct reasoning. We finally propose the SFT+RL approach with a **reasoning reward** to reinforce the generation of thinking tokens as a mental model abstraction, ultimately achieving a **significant 32.86% improvement**.
  - Our novel approach to reinforce thinking tokens via RL as an abstraction of maps is **new to all previous work**. We did not give a fancy name to this RL algorithm, but chose to write the paper in a way to bring audience along with us through the design space, which does NOT mean our method is trivial.
- We appreciate reviewer acknowledge "the contributions are primarily on the language side, with minimal novelty on the visual front", it is exactly scoped by our title "mental modeling", aiming to highlight this "abstraction" learning perspective. We totally agree vision front reasoning is challenging, but we **respectfully argue that language-centric reasoning is meaningful and needs better exploration, and cannot be simply viewed as non-exciting**.

2. **Scene Graph vs Map**

- We did not select scene graphs as an abstraction because JSON based maps have been widely adopted in the community [3,8]. Scene graphs focus on object status, while maps are about spatial abstraction. If we make scene graphs include object positions and orientations, it essentially becomes the map we adopted.
- New experiments show JSON consistently outperforms alternatives. It provides strong empirical evidence that our choice was **a principled decision, not merely one of convenience**.

Frozen VLM	Accuracy (%)
Map (JSON)	**41.33**
Natural Language Grounding	39.71
Symbolic Scene Graph (Plain)	37.90
Symbolic Scene Graph (JSON)	40.38
Vector-based formats	36.10

[1] What Makes a Maze Look Like a Maze?

[2] Retrieving Memory from a Cognitive Architecture by Impressions from Language Models for Use in a Social Robot

[3] Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces

[4] Perspective-Aware Reasoning in VLMs via Mental Imagery Simulation

[5] Hyperphantasia: A Benchmark for Evaluating the Mental Visualization Capabilities of Multimodal LLMs

[6] StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts

[7] LayoutLM: Pre-training of Text and Layout for Document Image Understanding

[8] SpatialLM: Training Large Language Models for Structured Indoor Modeling

Again, we thank you for valuable feedback, which helped us strengthen our paper. We hope our response fully address the concerns.

Official Comment by Reviewer WJos

Official Comment by Reviewer WJos 07 Aug 2025, 13:24 (modified: 07 Aug 2025, 18:20)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Revisions (/revisions?id=pP2Ept863e)

Comment:

> We respectfully doubt that the reviewer has a bias that language-based improvements are less important than visual-based.

The remark is not only presumptive, but borders on personal accusation. **This is not a fair or professional way to respond to legitimate scientific critique, respectfully.** My review raised concerns about the lack of connection to existing work on neither spatial language nor sparse view reconstruction, which is NOT a bias against language-based research.

It's not that I hold a bias against language-based contributions; rather, the way this paper is written gives the strong impression that the authors did not intend to engage meaningfully with the language side of spatial reasoning in the first place. If the claim to be contributing to “language-based spatial reasoning”, then why is there absolutely no reference in the related work section? There is an extensive literature on how spatial relations are expressed, inferred, and grounded in language (e.g., via landmarks, trajectors, prepositions, spatial roles). Research on space and language dates back to the HCRC Map Task in the 1990s. None of this is acknowledged. The paper treats spatial language as a downstream prompting issue rather than a deep modeling problem. So rather than addressing the substantive concern (i.e., the disconnect from prior cognitive and linguistic work), the rebuttal doubles down on an isolated VLM-centric framing of “spatial cognition”. This shows a lack of depth in understanding the broader research landscape the paper claims to contribute to.

Below are some concrete action items, that IMO, can meaningfully connect this work to the past work on spatial language and improve:

- Explicitly discuss past work on spatial language and spatial representations with language (e.g., spatial role labeling);
- Motivate the JSON design of Cognitive Map from previous work on structured representations of spatial relations.
- Adopt one of the past design as a baseline. I will use spatial role labeling as an example: the keys of the JSON-formatted Cognitive Map can actually be the "trajector", "landmark", "indicator" etc in the annotation scheme (currently it's augmented without justification). This will help the community understand how this work connects to the past, and how past work serves as a baseline for current modern VLM design.

> New experiments show JSON consistently outperforms alternatives. It provides strong empirical evidence that our choice was a principled decision, not merely one of convenience.

This addressed my concerns. I actually expect an experiment on YAML formats, but it's my problem for not being clear enough in previous discussions, so I am not requesting this to be added. Thanks.

Replying to Official Comment by Reviewer WJos

Official Comment by Reviewer WJos

Official Comment by Reviewer WJos 07 Aug 2025, 18:35

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Comment:

To build on the previous discussions, if the paper (1) is intended to and (2) can place greater emphasis on spatial language, with explicit connections (similar to the proposed experiment design I mentioned in the previous thread, but not necessarily follow what I proposed) and proper acknowledgments made, I would be willing to raise my score to borderline accept.

I believe the additional experiment can be completed before the camera-ready after the discussion period, and is much lighter compared to the commitments the authors have made to other reviewers. I acknowledge the value of this benchmark (this work is indeed more suitable for the benchmark track) and the solid efforts made during the

discussion phase.

Results on YAML formats

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

08 Aug 2025, 15:07 (modified: 08 Aug 2025, 20:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Revisions (/revisions?id=EUEC9cNdjB)

Comment:

We sincerely thank the reviewer for the insightful comments and for upholding the high standards of NeurIPS. The comments have greatly helped us to improve the paper.

Performance of using YAML format

We appreciate the suggestion of using YAML formats and have added the results below, which is consistent with our previous results. The potential reason is that current VLMs/LLMs have observed much training data of JSON formats and thus widely adopted in the previous work.

Representation	Accuracy (%)
Map (JSON)	41.33
Map (YAML)	40.38

Literature like Spatial Role Labeling

We greatly appreciate the reviewer's clarification regarding our previous questions. Our previous response was intended to better understand the feedback on balancing language and visual contributions, particularly given the statement "The contributions are primarily on the language side, with minimal novelty on the visual front. While the paper touches on an important and challenging problem, it feels like an early step rather than a deep exploration, and I find it difficult to get genuinely excited about the current scope of the work". We recognize this may not have been expressed as clearly as intended, and apologize for any misunderstanding. We are genuinely grateful for the reviewer's expertise in spatial language research prior to VLMs, which is valuable for strengthening our work's foundations.

The reviewer's suggestions align well with our research goals in the VLM context and will help readers better understand how our work connects to traditional foundations. We are excited about the opportunity to explicitly bridge our VLM-focused approach with the rich foundations prior to VLMs:

1. **Introduction:** We will explicitly position our contribution as linking spatial mental models to VLM reasoning through training paradigms that learn reasoning tokens, rather than relying on prompt-only tricks. We will make it clear that our tasks (Rotation, Around, and Among) probe relational reasoning across viewpoints and perspective-taking, which aligns with figure/ground and landmark-anchored behaviors discussed in spatial language. Furthermore, we will add an orienting paragraph that acknowledges the HCRC Map Task, Spatial Role Labeling, frames of reference (relative/intrinsic/absolute), and motion semantics (figure–ground, path/manner) as the conceptual background for our structured representation choices.
2. **Method:** We will motivate the adoption of JSON as a transparent, verifiable, and compositional intermediate representation that supports schema checks and graph-level consistency metrics (e.g., validity, overall similarity, isomorphism) and allows the model to first construct a map and then reason over it, mirroring our evaluation pipeline. To justify our schema, we will connect our Cognitive Map design to Spatial Role Labeling by explicitly aligning role keys with first-class JSON fields: `trajector`, `landmark`, `indicator` (preposition), `frame_of_reference`, and `direction`, while preserving our existing `views` (`[{position, facing}]`) for viewpoint coherence.
3. **Results & Discussion:** We will add representation ablations to the Appendix (JSON cognitive map vs. YAML vs. scene-graph formats) under identical settings to the reported scaffolds for comparability, and include a concise discussion showing that map-then-reason together with an SFT→RL training paradigm that induces role schemas yields the largest gains.
4. **Related Work:** We will merge the current “Spatial Cognition” section with spatial-language work into a subsection placed after the VLM/benchmark paragraph for an ML→linguistics flow, and close by explaining how these roles organize our analyses (by preposition class, frame of reference, and path) and how our role-aligned map differs from prior object-centric “cognitive maps”.

We appreciate the reviewer's constructive engagement and look forward to strengthening the connection of our VLM-focused approach and traditional spatial language literature prior to VLMs.

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

08 Aug 2025, 15:21 (modified: 08 Aug 2025, 18:38)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer WJos

Revisions (/revisions?id=HXo2gc5cod)

Comment:

Also, we hope to take the opportunity to detail the motivation of our current related work organization, which is center around the emerging field of Spatial Intelligence for VLMs to position our contributions within this rapidly evolving research area.

Our related work section focuses primarily on VLM approaches after 2023 because we aimed to highlight how the VLM spatial intelligence field has rapidly progressed from basic single-view reasoning to complex multi-view understanding. We are genuinely excited about the a well-recognized challenge that although modern VLMs after GPT4-V have improved on general reasoning impressively, but not on spatial reasoning. The field initially focused on static, single-view reasoning, like VSR [30] and Whats UP [18] focusing on simplistic, atomic spatial prepositions like 'left of' or 'above' from a single image. Recognizing these limitations, the community progressed to multi-view benchmarks which leverage multiple images or short video clips, like SPAR-Bench [31] on low-level perception to high-level reasoning, 3DMV-VQA [32] on 3D concept reasoning from actively gathered views, STI-Bench [12] on spatio-temporal dynamics, SpatialScore [17] on fine-grained visual geometry perception of VLMs, VSI-Bench [33] on object size, position, and relative relationships from both qualitative and quantitative perspectives. Typical solutions like SpatialRGPT [36] and SpatialBot [34] use depth encoders to provide rudimentary geometric cues alongside RGB input.

Compared to these related work in spatial intelligence of VLMs, our investigation led to several surprising findings that challenge the conventional wisdom in the community, moving to build and maintain a spatially consistent mental model during interactions, so we emphasize "spatial cognition" and "multi-view understanding" in our related work, and contextualize both of them in VLM research. For example, we discovered that **teaching the model how to think with an abstraction is far more effective than just giving it more to see**. We believe uncovering such non-obvious phenomena and showing how to use it with a impressive 32.86% absolute score improvement, can point to future research in spatial intelligence for VLMs to perform complex spatial tasks: an explicit and structured internal representation is not just an optional add-on, but a **core prerequisite for robust spatial reasoning** in VLMs.

The reviewer's suggestions will strengthen our work greatly by contextualizing our contribution not only in emerging VLM research, but also in the traditional spatial language domain. It will demonstrate how classical spatial language insights can inform and validate emerging VLM spatial reasoning approaches, by explicitly connecting these VLM advances to the theoretical foundations prior to modern VLMs. **Again, we sincerely appreciate the reviewer's suggestions and detailed feedback.**

Official Review of Submission4015 by Reviewer 89VQ

Official Review by Reviewer 89VQ 30 Jun 2025, 22:05 (modified: 18 Sept 2025, 11:31)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Revisions (/revisions?id=OjoNhXtzBn)

Summary:

This paper introduces the MindCube benchmark, which aims at evaluating the spatial reasoning capability of large VLMs when limited views are available. Using the proposed benchmark, the authors conduct a systematic analysis of the spatial reasoning capabilities of VLMs under multiple views, including direct prompting, fine-tuning, and reinforcement learning. These experiments highlight the challenges and limitations of current VLMs in this area, as well as suggesting promising post-training strategies (SFT followed by RL) as a practical solution.

Strengths And Weaknesses:

In what follows, I will use S, W, C, and Q to denote Strengths, Weaknesses, Comments, and Questions, respectively, for ease of reference in the discussion.

S1. The idea of spatial reasoning with partial observations is an important and underexplored aspect in VLMs. The benchmark is fairly comprehensive, covering three camera-movement patterns and formulating the task into multiple-choice question-answering problems. I anticipate that the benchmark will enable a line of work on spatial reasoning with VLMs, particularly highlighting the importance of (limited) multi-view based reconstruction.

S2. The authors have conducted a series of experiments (including direct prompting, fine-tuning, and reinforcement learning) that show the effectiveness and limitations of current VLMs. Even with the strongest fine-tuning + RL method, the model performance is still far from perfect, which indicates that the benchmark is challenging.

S3. The experiment section establishes a "standard" approach to evaluate the spatial reasoning capability of VLMs, which I believe will be useful for future work in this area.

W1. I understand that this is probably not a widely adopted approach in AI, but this paper will benefit a lot from reporting statistical significance of the results, especially when performances are fairly close (e.g., Table 3). Having error bars in Figure 4 wouldn't be too hard, but would add a lot to the clarity and confidence of the results. Testing statistical significance over a few variations of prompts would also help.

W2. Beyond the accuracy-driven evaluation and the current failure case analysis (Appendix B.2), which looks like randomly picking two failure cases, I would like to see a more systematic analysis of the model's failure cases. For example, does something like self-consistency help? If the model fails, does it usually show some failure patterns? I believe a systematic taxonomy of failure analysis would be very helpful for future work that follows the benchmark.

W3 (C1). I'm not sure if this should be considered as a weakness under the current intellectual climate, since many papers with similar or even greater limitations have been published in our top venues, and it may be unreasonable to single out this work for such criticism. Nonetheless, the investigated approach (prompting, SFT, RL) is unlikely to be the ultimate solution due to a lack of interpretability and principled design choices beyond existing VLMs and alignment techniques. Please see more detailed comments below for some more principled examples that I believe are worth referring to.

W4 (C2). This is again unnecessarily a weakness---I understand this may be because of the time/space limit, but a few experiments could be made stronger by tuning some hyperparameters and the reward function.

Some more details comments

C1. While I'm trying not to be biased against the presented prompting approach, I would like to mention that simply prompting/finetuning VLMs with predetermined architectures might not be the best way (or, unlikely to be an ultimate solution, in my personal and possibly biased opinion) of building a comprehensive world model due to lack of interpretability.

This is not a citation requirement---since the papers I am sharing were released fairly close to or even after the NeurIPS deadline, and they are not directly related in terms of methodology---I believe it is worth mentioning in case the authors are not aware of this line of work: https://arxiv.org/pdf/2505.00702 (https://arxiv.org/pdf/2505.00702), https://arxiv.org/pdf/2506.18890 (https://arxiv.org/pdf/2506.18890).

C2. I would like to see more systematic hyperparameter tuning and reward function design in the RL experiments. For example, the authors could try different reward functions and see if there's a difference in performance---the current +1 and +5 look somewhat arbitrary, though reasonable.

C3. Beyond performance (behavioral) evaluation, some competence-driven analysis of the model's spatial reasoning capabilities would strengthen this paper a lot. For example, these could be some interesting approaches to explore:

1. Taking the intermediate hidden states (i.e., activations) of VLMs and analyzing them with some standard probing approach.
2. Would there be some "shared neurons" across the views that represent the same object?
3. Since the authors also annotated ground-truth cognitive maps (L62-63), it would be interesting to hold out a small portion and see if there are systematic similarities and dissimilarities between the generated cognitive maps and the ground truth.

C4. Writing

1. There are a few words that I would like to suggest changing: for example, *soar* (L268) is somewhat uncommon in academic writing, and I would suggest using something like *achieve* or *rise*, or even *surge* instead.

   Also, the expression *RL shines* sounds like an advertisement for RL, and I would suggest replacing it with more neutral and objective expressions like *RL performs well/better*.

   I also believe the usage of *sharp* (L232, L245) is a bit off---it is unnecessarily anthropomorphic, and it's particularly worth noting that "sharp" is usually used in another sense in machine-learning literature (e.g., sharp bounds)---I would suggest using *comprehensive* or *competent* instead.
2. L292: Multi Views understanding -> Multi-View Understanding?
3. It would be good to unify the decimal point notation throughout the paper. For example, I had difficulty allocating the 38.3 and 76.1 referred to in the abstract and introduction.

Quality: 2: fair
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:
I am a bit confused about the poor performance of CGMap-In-FFR-Out (L186) and not convinced by the authors' hypothesis, and the question below are all related to this point.

Q1. How did you specify which image is Image 1/2/3/4 to the model?

Q2. (Related to C3) have you compared the generative cognitive map with the ground truth?

Q3. How about the performance of "still outputting a cognitive map like Plain-CGMap-FFR-Out, even if the input is given", which I believe is not tested in the current paper?

Limitations:

The authors claim that they have discussed the limitations in the checklist; however, it is more of the limitations in current VLM-based spatial reasoning approaches instead of the limitations of this work itself. Apparently, this work is limited in terms of finer-grained analysis and interpretation; however, unfortunately, the authors didn't discuss their own work's limitations properly. I encourage the authors to consider drafting a limitation section in rebuttal.

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Ethical Concerns: NO or VERY MINOR ethics concerns only

Paper Formatting Concerns:

N/A

Code Of Conduct Acknowledgement: Yes

Responsible Reviewing Acknowledgement: Yes

Final Justification:

Sorry, I didn't realize this was a mandatory update initially. As I indicated in comments and the reviewer-AC discussion, I'm not in support of this paper due to the intensive new results offered in rebuttal, which serve as an indicator that the paper needs another round of review (also pointed out by Fellow Reviewer dbE7).

Rebuttal by Authors

Rebuttal

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

31 Jul 2025, 07:44 (modified: 31 Jul 2025, 14:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=nBhQZJbMGC)

Rebuttal:

We sincerely thank reviewer for the constructive feedback. We are grateful that the reviewer finds our proposed benchmark comprehensive and challenging, the topic important and underexplored, and our experimental approach a useful "standard" for future work.

The concerns raised are valuable, and we believe our new experiments and clarifications below fully address them.

1. **(W1)** We agree and have conducted new experiments with multiple random seeds to provide statistical analysis.
   - We have re-run our key experiments with multiple seeds (3 runs) to report the mean and standard deviation, and below are the results:

Frozen VLM Config.	**Acc**
Raw-QA	36.19±5.95
Aug-CGMap-In	33.78±0.73
FF-Rsn	40.35±0.83
Aug-CGMap-FFR-Out	37.87±5.93
Plain-CGMap-FFR-Out	38.22±2.67
CGMap-In-FFR-Out	37.59±0.67

   - We are curious about SFT configs when doing multiple runs, and the results (below table) **still align with our main claim that "map-then-reason" is the most effective approach**.

SFT Config.	**Acc**
Raw-QA	51.14±0.90
FF-Rsn	51.27±1.53
Aug-CGMap-Out	51.14±1.16
Plain-CGMap-Out	52.35±1.67
Aug-CGMap-FFR-Out	**53.28±1.34**
Plain-CGMap-FFR-Out	**56.79±1.06**

- **Prompt Sensitivity Analysis.** As suggested, we analyzed prompt sensitivity by comparing our structured representation (S) against a natural language (NL) paragraph.

Setting	Acc
Aug-CGMap-FFR-Out (S)	40.57
Aug-CGMap-FFR-Out (NL)	37.43
Plain-CGMap-FFR-Out (S)	41.33
Plain-CGMap-FFR-Out (NL)	39.71
CGMap-in-FFR-Out (S)	41.43
CGMap-in-FFR-Out (NL)	43.43

  - These results indicate that the best format is setting-dependent.
  - This aligns with the principle that VLMs should adapt to varied user inputs, rather than users needing to find a perfect prompt format [1].
  - This is why our paper reports results from one consistent and representative prompt variation.

2. **(W2)** We agree that a systematic failure analysis is very valuable, and see if self-consistency (SC) can help.
   - We sampled 100 GPT-4o responses with reasoning and categorize the observed errors into four types:
   - Situation-Transformation Reasoning Error. The model misidentifies or mislocates objects in the image
   - Direction-Transformation Error. The model fails to update left/right to front/back after its viewpoint rotates
   - Viewpoint Error. The model misjudges its own movement (e.g., confuses a 90° turn with a 180° turn)
   - Grounding Error. The model fails to correctly identify or locate objects in the image
   - We implemented an automated error analysis pipeline using GPT-4o, and below are the results:

Error Type	Situation-Transformation	Direction-Transformation	Viewpoint	Grounding
ratio	36.00	13.00	48.00	3.00

   - Key Findings
     1. A model's reasoning process can be flawed even when its final answer is correct.
     2. Among all error types, **"Situation-Transformation Error"** and **"Viewpoint Error"** are the most common.
   - **SC barely helps.** Below are our results:

Frozen VLM Config.	Overall
Raw-QA	36.29 (↓1.52)
Aug-CGMap-In	32.95 (↓0.95)
FF-Rsn	40.57 (↑0.07)
Aug-CGMap-FFR-Out	31.05 (↓9.52)
Plain-CGMap-FFR-Out	35.62 (↓5.71)
CGMap-In-FFR-Out	37.71 (↓3.72)

3. **(W3 & C1)** We thank the reviewer for their thoughtful perspective on the broader landscape of world modeling.
   - **Our scope is to benchmark and find effective representations for VLMs to do spatial reasoning, a distinct goal from designing new, interpretable architectures.**
   - We agree that current alignment techniques are likely not the final solution for building comprehensive world models.
     - However, our paper's central question is how to best approximate spatial mental models within current VLM paradigms.
     - **We do not claim our method builds a complete world model, but that it ("map-then-reason", the representation) enhances a critical component of spatial intelligence.**
     - We also agree that finding effective representations for current architectures is a crucial and complementary research direction to developing new ones.
4. **(W4 & C2)** We thank the reviewer for these constructive suggestions and have conducted new hyperparameter and reward function tuning experiments.
   - We conducted hyperparameter tuning for our SFT experiments in `Raw-QA` setting.
   - Our results confirm that the hyperparameters used in our submission are effective.

SFT (learning rate, batch size, warmup ratio)	Acc
(1e-5, 512, 0.03) -- Ours	**52.28**
(2e-5, 512, 0.03)	51.71
(4e-5, 512, 0.03)	51.52

SFT (learning rate, batch size, warmup ratio)	Acc
(1e-5, 256, 0.03)	50.86
(1e-5, 1024, 0.03)	51.90
(1e-5, 512, 0.01)	51.81
(1e-5, 512, 0.10)	50.67

- We also systematically investigated the RL reward function design for the `FF-Rsn` setting as suggested.

Reward (format score, answer score)	Acc
Frozen VLM (baseline)	40.57
Our Setting (1, 5)	50.57
Answer Only (0, 5)	50.76
Format Only (5, 0)	45.05

  - Rewarding only for the correct answer yields a comparable result to our setting.
  - Removing the reward for the correct answer significantly degrades performance.
  - This confirms that task correctness is the most critical reward component, consistent with recent work [2].

5. **(C3)** We thank the reviewer for these insightful comments.
   - To test for a viewpoint-invariant object representation, we analyzed 10th-layer LLM activations for 100 object triplets (front/left/right views) from MindCube.
     1. Pairwise Similarity: We computed cosine and Pearson similarities on activations for same-object pairs across views (positive) versus different-object pairs (negative). In layer 10, similarities for positive pairs were significantly higher than for negative pairs, indicating a viewpoint-invariant representation of object identity.

Metric	Positive pairs	Negative pairs
Cosine	0.9651	0.9160
Pearson *r*	0.2750	0.2046

     2. Stable-Dimension Search: We identified 64 low-variance "stable dimensions" across the three views. Within these dimensions, activations for the same object showed extremely high cosine similarity (0.931 ± 0.068), confirming they act as viewpoint-invariant neurons.
     3. Probing: A linear probe trained on activations from two views successfully predicted the 90° rotation direction for unseen objects with 85% accuracy (chance: 50%), confirming the representation generalizes.
   - We qualitatively analyzed model-generated maps against the ground truth, focusing on typical failures like object misplacement and errors in viewpoint or object orientation.
   - Due to space constraints here, we will add visualizations of cogmap comparison visualization to the final version paper.
6. **(C4)** We sincerelly appreciate the detailed and constructive suggestions to improve the manuscript's clarity and professionalism and have revised the text accordingly in our manuscript.
7. **(Q1) Image Order Specification.**
   - Image order is handled both implicitly by the VLM architecture and explicitly in our prompts.
   - The VLM architecture inherently processes inputs sequentially and uses distinct special tokens to represent each image in the order it is provided.
   - We also explicitly specify image IDs in our questions (e.g., "Based on the input images (image 1, 2, 3, 4)..." "From the viewpoint presented in image 2").
8. **(Q2) Generated Cognitive Maps Evaluation.**
   - **Yes, we quantitatively evaluate all generated maps against ground-truth maps.**
   - We propose a suite of graph-based metrics (Valid Rate, Isomorphic Rate, Overall Similarity) for this evaluation, defined in Appendix C.2 of our submission.
9. **(Q3) Input Given `CGMap-FFR-Out`.** We thank the reviewer for this excellent suggestion and found this really insightful.
   - **We ran this experiment; the results reinforce our core claim that reasoning over a provided map is highly effective.**
   - The table below presents the full updated results:

Frozen VLM Config.	Overall
CGMap-In-FFR-Out	41.43
CGMap-In-CGMap-Out (New)	26.86

Frozen VLM Config.	Overall
CGMap-In-CGMap-FFR-Out (New)	**45.90**

- Regenerating a map when one is already provided (`CGMap-In-CGMap-Out`) significantly degrades performance.
- However, first regenerating the map and then reason (`CGMap-In-CGMap-FFR-Out`) performs the best among all frozen VLM settings.
- These new results reinforce a key insight of our paper: the "map-then-reason" paradigm is an effective cognitive process, especially when grounded with an explicit map as input.

[1] *Holistic Evaluation of Language Models.* Percy Liang, et al.

[2] *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* DeepSeek-AI.

Limitations

Our reasoning approach centers primarily on language-based question-answering formats. Second, we assume idealized input by using images without noise. Third, our dataset shows class imbalance across different categories. In the future, we will extend to spatial reasoning beyond language-centric thinking. Due to the space limit, we will include a complete version in the revision.

Replying to Rebuttal by Authors

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

03 Aug 2025, 12:08 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewer 89VQ,

Thank you again for your thoughtful and constructive review, and we would like to kindly follow up on your opinions on our rebuttal.

We worked extensively to address the points you raised, and we are keen to ensure our revisions are satisfactory and would be grateful for your feedback when you have a moment.

Thank you for your time and consideration.

The Authors

Thank you for your additional experiments; ambivalent about the paper.

Official Comment by Reviewer 89VQ 05 Aug 2025, 15:47 (modified: 05 Aug 2025, 15:47)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=Fp12BeBh0j)

Comment:

I would like to thank the authors for the additional experiments, and this would definitely make the paper stronger on the technical end.

On the one hand, I think the paper in its current shape is fairly complete; however, I am ambivalent about the high-level pictures, as the paper does not much deepen our understanding of the spatial reasoning task, and the failure care analysis, as well as the new results, probably needs another round of review to ensure scientific rigor. I would have been willing to raise my rating to something like 3.5, but in calibration with Reviewer dbE7, I'll keep my rating for now.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer 89VQ

Mandatory Acknowledgement by Reviewer 89VQ 05 Aug 2025, 15:47

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Can the reviewer provide more clarification about the concerns?

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

06 Aug 2025, 22:40 (modified: 07 Aug 2025, 07:30)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Revisions (/revisions?id=XT51ehVyuW)

Comment:

We sincerely thank the reviewer for their constructive feedback. However, it would be greatly appreciated if the reviewer could provide further clarification on the remaining concerns.

1. **Need further clarification about reviewer concerns on "deepening the understanding of spatial reasoning"**

- We are confused about why current paper is not deepening the understanding of spatial reasoning, since we **exhaustively explored the design space of language-based abstraction** with 12 design choices. From frozen models to different SFT and RL paradigms, we finally propose the SFT+RL approach with a **reasoning reward to reinforce the generation of thinking tokens as a mental model abstraction** (which is new), ultimately achieving a **significant 32.86% improvement** over strong baselines.
- **“High-Level Picture”**: We deepen understanding of spatial reasoning by answering **important questions** in VLMs for spatial reasoning:
  - "What is different between spatial reasoning vs general reasoning in large reasoning models?"
  - "What granularity of abstraction works for VLMs?"
  - "RL cannot directly help much. What kind of reward design is needed?"
  - "When is RL ready to help VLMs emerge spatial intelligence? Is a warm-up/cold start needed before RL? What kind of cold start is most effective, showing what kind of info is needed as prior?”
- We conduct **systematic exploration** to answer these questions:
  - We collected **17,530 questions** covering 3 view changes, 4 what-if dynamics, 3 types of relation queries, 2 types of perspective taking, two types of visual patterns.
  - Even latest model get **close-to-random performance** in our data.
  - Our novel approach to reinforce thinking tokens automatically as an abstraction of maps via RL, is **new to previous work** [1, 2, 3]. We did not give a fancy name to this RL algorithm, but chose to write the paper in a way to bring audience with us along exploration of the design space, which does NOT mean our method is not new.
- To the best of our understanding, we respectfully doubt that the reviewer has **a bias that language-based reasoning are less important than visual-based**, which is a potential reason why the reviewer feels there is not much deepen understanding of spatial reasoning.
  - For decades, spatial reasoning has been a problem from both vision and language perspectives. The language perspective has also been well recognized as a line of research.
  - Simply offering denser visual info is insufficient for spatial reasoning. It has been empirically proved in Lines 51-53 and Table 3 where denser visual info did not lead to performance gains.
  - **Language and Vision are simply different perspectives for advancing spatial reasoning. It is beneficial to encourage both communities to contribute to this goal, rather than diminishing the insights emerging from the text domain.**

2. **Need further clarification about concerns on scientific rigor**

- We thank the reviewer for their focus on scientific rigor and **hope the reviewer could point out which result is a concern for scientific rigor**. It is irresponsible to vaguely say more review is needed for scientific rigor but without specific concerns about which results are concerning. We would love to address any further concerns regarding the scientific rigor if the reviewer can be more specific.

- Our experiments were conducted as a **strict, controlled comparison**: when testing different data scaffolds (e.g., cognitive maps vs free-form reasoning), the base model, dataset, and evaluation protocol were all held constant. The resulting conclusions are therefore robust.
- Per the reviewer's request, we show **10 latest experiment and error analysis tables** to address potential concerns, also conducted in a **rigorous, controlled way**.
- For example, we include 7 different hyperparameter searching experiments for SFT and 4 for RL.
- For error analysis, we sample error cases and do analysis by following prior works [1, 2, 3] in this field.
- To ensure the analysis quality, we implement both an automated (GPT-4o) pipeline and manual checking. If there is disagreement, a third person will make the final decision.
- To be fully transparent, we **have included a detailed 40-page appendix with all experimental settings, hyperparameters, and will add all results in the rebuttal as expanded analysis in the revision**.
- The code are attached in Supplementary and will be publicly available.
- We are confident this commitment demonstrates the scientific rigor of our work for the current review cycle.

[1] Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces.

[2] Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs.

[3] Perspective-Aware Reasoning in VLMs via Mental Imagery Simulation.

Again, we thank the reviewer for valuable feedback, which has helped us strengthen our paper. We would love to address any further concerns.

Replying to Can the reviewer provide more clarification about the concerns?

Official Comment by Reviewer 89VQ

Official Comment by Reviewer 89VQ 07 Aug 2025, 09:25

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Comment:

I think we are already on the same page about this paper's technical contribution but just not on what scientific rigor is as well as valuable work. For work that I value as significantly deepening our understanding of something, I'd look for something that fundamentally **solves** the problem with guarantees or proposes a fundamentally challenging problem that we've not been aware of, but not applying some existing methods that we don't understand much why they work, just like many others have done in the community; otherwise, it's just okay to either publish or not, and probably not at NeurIPS, and given that we're already having too many such papers, I would be cautious to recommend another one.

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

08 Aug 2025, 18:29 (modified: 09 Aug 2025, 08:00)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Revisions (/revisions?id=kiw3fTmaba)

Comment:

We sincerely thank the reviewer for the insightful comments and for upholding the high standards of NeurIPS. We agree with the reviewer's emphasis on work that identifies previously unrecognized challenges or fundamentally solves problems. Our work addresses such a fundamental challenge: the missing abstraction interface between visual perception and high-level reasoning in VLMs. Such abstraction is a cognitive mental model.

While VLMs have achieved remarkable progress in both perception and high-level reasoning independently, spatial reasoning performance remains a significant bottleneck. Our investigation reveals that this bottleneck stems from a fundamental architectural limitation: the lack of maintaining spatially consistent mental models as an abstraction between perception and high-level reasoning, so that VLMs can predict unobservable scenes even with partial observations.

Our work resides in the emerging and challenging field of Spatial Intelligence for VLMs. To contextualize our contribution, we wish to highlight the current consensus within the Spatial Intelligence research community. Prior work in spatial reasoning for VLMs is primarily in clean, unambiguous scenes.

The field initially addressed static, atomic spatial relations like VSR [30] and What's Up [18], which handle simple prepositions such as 'left of' or 'above' from single images. The field initially focused on static, single-view reasoning, like VSR [30] and Whats UP [18] focusing on simplistic, atomic spatial prepositions like 'left of' or 'above' from a single

image. Recognizing these limitations, the community progressed to multi-view benchmarks which leverage multiple images or short video clips, like SPAR-Bench [31] on low-level perception to high-level reasoning, 3DMV-VQA [32] on 3D concept reasoning from actively gathered views, STI-Bench [12] on spatio-temporal dynamics, SpatialScore [17] on fine-grained visual geometry perception of VLMs, VSI-Bench [33] on object size, position, and relative relationships from both qualitative and quantitative perspectives. Typical solutions like SpatialRGPT [36] and SpatialBot [34] use depth encoders to provide rudimentary geometric cues alongside RGB input. While this was a large step forward in the past two years, but they typically treat these powerful visual-front encoders as frozen, black-box feature extractors, creating a disconnect between the representation and the final reasoning process. This highlights a significant architectural limitation in the current state-of-the-art: We may need better abstraction to connect and perception and high-level reasoning. Our work represents a novel attempt to address this challenge.

Crucially, our investigation led to several surprising findings that challenge the conventional wisdom in the community. For example, we discovered that teaching the model how to think (map-then-reason) is far more effective than just giving it more to see. This insight opens up new avenues for future research in spatial intelligence. The significant 32.86% improvement suggests that for complex spatial tasks, an explicit and structured internal representation is **not just an optional add-on**, but a **core prerequisite for robust spatial reasoning**. We believe uncovering such non-obvious phenomena is precisely the kind of work that "deepens our understanding".

We appreciate the reviewer tends to raise the score to 3.5 and we will make these comparisions and the additional results in the rebuttal included in the revision.

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

09 Aug 2025, 07:58 (modified: 09 Aug 2025, 08:19)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Revisions (/revisions?id=gqG0gqQq5y)

Comment:

Thank you again for your valuable time and insightful feedback.

To provide a clear and convenient overview of the current research landscape, we have compiled the detailed comparison table below. Many of these works, including several very recent ones, are already cited within our manuscript. However, to help the reviewer intuitively grasp the field's evolution and our work's positioning, we are presenting them here together, and will list their full details at the end of this response.

As the table illustrates, the field has made significant progress, yet existing benchmarks often focus on a subset of spatial skills. Our work is designed to bridge the remaining gaps by introducing not just a new benchmark, but a **comprehensive framework** centered on a core methodology.

Our analysis suggests that higher-order cognitive abilities like **Free-Form Reasoning (FFR)**, **Perspective-Taking**, and **Consistency** are crucial for advancing beyond simple spatial perception. Our **MindCube framework** is distinctive in its holistic integration of these diverse challenges. The design of our benchmark is novel and comprehensive in that it synthesizes:

- **Higher-Order Cognition**: A combined focus on `FFR`, `Perspective-Taking`, and `Spatial Consistency`.
- **Complex Perception**: Support for `Multi-view` inputs with controlled `Camera Layouts`.
- **Environmental Diversity**: Coverage of both `Indoor` and `Outdoor` scenes.

While our benchmark does not currently focus on fine-grained Quantitative questions, this was a deliberate design choice. Our priority is to address the more fundamental and pressing challenge of building a coherent mental model, rather than precise numerical calculation.

This integrated approach—encompassing both our methodology and the resulting benchmark—is designed to rigorously test for the "spatial mental model" and push the field towards models that can form and reason over spatially consistent mental models, thus clarifying the novel and significant contribution of our work.

Thank you once more for your detailed review and guidance.

References from the Comparison Table

[1] What's "up" with vision-language models? Investigating their struggle with spatial reasoning: https://arxiv.org/abs/2310.19785 (https://arxiv.org/abs/2310.19785)

[2] Visual Spatial Reasoning: https://arxiv.org/abs/2205.00363 (https://arxiv.org/abs/2205.00363)

[3] Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs: https://arxiv.org/abs/2406.16860 (https://arxiv.org/abs/2406.16860)

[4] SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models：https://arxiv.org/abs/2406.01584 (https://arxiv.org/abs/2406.01584)

[5] SpatialBot: Precise Spatial Understanding with Vision Language Models：https://arxiv.org/abs/2406.13642 (https://arxiv.org/abs/2406.13642)

[6] SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models：https://arxiv.org/abs/2412.07755 (https://arxiv.org/abs/2412.07755)

[7] Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces：https://arxiv.org/abs/2412.14171 (https://arxiv.org/abs/2412.14171)

[8] RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics：https://arxiv.org/abs/2411.16537 (https://arxiv.org/abs/2411.16537)

[9] EmbSpatial-Bench: Benchmarking Spatial Understanding for Embodied Tasks with Large Vision-Language Models：https://arxiv.org/abs/2405.09349 (https://arxiv.org/abs/2405.09349)

[10] 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark：https://arxiv.org/abs/2412.07825 (https://arxiv.org/abs/2412.07825)

[11] MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs：https://arxiv.org/abs/2503.13111 (https://arxiv.org/abs/2503.13111)

[12] STI-Bench: Are MLLMs Ready for Precise Spatial-Temporal World Understanding?：https://arxiv.org/abs/2503.23765 (https://arxiv.org/abs/2503.23765)

[13] OmniSpatial: Towards Comprehensive Spatial Reasoning Benchmark for Vision Language Models：https://arxiv.org/abs/2506.03135 (https://arxiv.org/abs/2506.03135)

[14] SpaCE-10: A Comprehensive Benchmark for Multimodal Large Language Models in Compositional Spatial Intelligence：https://arxiv.org/abs/2506.07966 (https://arxiv.org/abs/2506.07966)

[15] MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence：https://arxiv.org/abs/2505.23764 (https://arxiv.org/abs/2505.23764)

Official Comment by Authors

Official Comment

by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

09 Aug 2025, 07:58 (modified: 09 Aug 2025, 08:16)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 89VQ

Revisions (/revisions?id=dYoeBP16Sz)

Comment:

Legend: ✓ = Supported/Evaluated | × = Not Supported/Evaluated | R = Real-world | S = Simulated | - = Not Applicable | Camera Layout = Whether camera positions have specific spatial distribution in multi-view | FFR = Free Form Reasoning | Consistency = Spatial Consistency Perception from multiview images | Spatial Orientation = Awareness of Object Orientation

Benchmark	QA Pairs	Multi-view	Env.	Camera Layout	Indoor	Outdoor	Quantitative	Spatial Orientation	FFR	Perspective-Taking	Consistency
What's up[1]	5K	×	R	-	✓	×	×	×	×	×	×
VSR[2]	10K	×	R	×	✓	✓	×	×	×	×	×
CV-Bench[3]	2.6k	×	R	×	✓	✓	×	×	×	×	×
SpatialRGPT-Bench[4]	1.4K	×	R	×	✓	✓	×	×	×	×	×
SpatialBot-Bench[5]	200	×	R	-	✓	✓	✓	×	×	×	×
SAT[6]	218k	✓	S	✓	✓	×	×	✓	×	✓	×
VSI-Bench[7]	5.1K	✓	R	×	✓	×	✓	✓	×	×	×
RoboSpatial[8]	1M	×	R	-	✓	×	×	×	×	×	×
EmbSpatial-Bench[9]	3.6k	×	R	×	✓	✓	×	×	×	✓	×

Benchmark	QA Pairs	Multi-view	Env.	Camera Layout	Indoor	Outdoor	Quantitative	Spatial Orientation	FFR	Perspective -Taking	Consistency
3DSRBench[10]	2.8k	✓	R	×	✓	×	×	✓	×	✓	×
Spatial-MM[11]	2.3K	✓	R/S	-	✓	×	✓	×	×	×	×
STI-Bench[12]	2.1k	✓	R	×	✓	×	×	✓	×	×	×
OmniSpatial[13]	1.5k	×	R/S	-	✓	✓	×	✓	×	✓	×
SpaceE-10[14]	6k	×	S	×	✓	×	×	×	×	×	×
MMSI-Bench[15]	1k	✓	R	✓	✓	×	×	×	✓	×	✓
MindCube (Ours)	**20k**	**✓**	**R**	**✓**	**✓**	**✓**	**×**	**✓**	**✓**	**✓**	**✓**

Official Review of Submission4015 by Reviewer 5ZSD

Official Review by Reviewer 5ZSD 25 Jun 2025, 03:58 (modified: 24 Jul 2025, 07:41)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 5ZSD

Revisions (/revisions?id=4b9WTZt5As)

Summary:
The paper tackles question of whether vision-language models (VLMs) can build human-like spatial mental models from a handful of egocentric photos. To probe this gap, the authors release MINDCUBE, a 2 919-image, 17 530-QA benchmark spanning three camera-motion regimes (Rotation, Around, Among) that force a model to infer the locations of occluded objects, adopt new viewpoints and answer "what-if" movement queries. Testing 14 leading VLMs shows that accuracy barely rises above chance —49 % at best, versus human performance near 95 %—confirming that existing systems lack robust spatial reasoning.

The study then asks how to scaffold that reasoning. Simple cues such as inserting interpolated frames or feeding the model an externally supplied bird's-eye map yield negligible gains. In contrast, prompting a frozen VLM (Qwen2.5-VL-3B) to first generate its own coarse cognitive map and then articulate a chain-of-thought lifts accuracy from 38 % to 43 %. Building on this idea, the authors curate 10 k cognitive-map labels and 10 k grounded reasoning chains for supervised fine-tuning. Fine-tuning on QA alone reaches 46 %, on reasoning alone 51 %, but combining map generation with reasoning drives performance to 61.7 % while tripling map-to-ground-truth isomorphism.

Finally, RL with GRPO is applied. RL from scratch scarcely helps, yet when warm-started from the best fine-tuned checkpoint it propels accuracy to 76.1 %, with map similarity surpassing 84 %. The authors conclude that the critical missing ingredient is not more views or ground-truth geometry; it is the act of making the model construct and reason over its own internal map.

Contributions are therefore: (i) the MINDCUBE benchmark, (ii) a controlled analysis of three scaffold types, (iii) evidence that self-generated cognitive maps plus free-form reasoning unlock the largest gains, and (iv) an RL recipe that sets a new state of the art for limited-view spatial reasoning.

Strengths And Weaknesses:
S1. Very nicely structured and maintained codes.

S2. Easy-to-understand illustrations and careful example selections.

S3. Systematic benchmark conceptual framework that could be reused for future VLM mental model and cognitive map studies.

S4. Showed surprising effectiveness of GRPO in enhancing models' spatial reasoning capability. Also provided good evidence that prior matters for LLM RL (i.e. need to do finetuning first, otherwise won't work so well).

W1. Toy-map abstraction – The model is asked to output a 10 × 10 discrete bird's-eye grid. Such a schematic ignores metric scale, elevation and real-world sensor noise, so the learned policy cannot be transferred to SLAM pipelines, indoor localisation or autonomous-driving coordinates where continuous 3-D poses and camera intrinsics matter.

W2. Idealised visual input – Training and evaluation use perfectly aligned 90° view hops and sharp RGB frames. Real embodied agents contend with blur, rolling-shutter artefacts, non-orthogonal rotations and missing frames; performance is therefore likely to collapse outside these pristine conditions.

W3. Static, template-driven QA on a single distribution – The task reduces spatial reasoning to multiple-choice questions with fixed distractor templates and no scene dynamics. Because the same template generator provides SFT labels and RL reward, the model may overfit shallow linguistic patterns rather than learn a general spatial representation; its ability to guide real-time interaction, handle moving objects or adapt to out-of-distribution language remains untested.

W4. GPT-O3 and O4-mini (high) has been available since the early 2025, and represents a significant breakthrough in visual perception by reasoning with images in their chain of thought. Even if it might not be a fair comparison to the smaller baseline models, the performance statistics are expected to be reported for a benchmark paper.
Quality: 3: good
Clarity: 4: excellent
Significance: 3: good
Originality: 2: fair
Questions:

1. What are the human performance on the benchmark? Is there a small-scale study?
2. What is the inter-annotator agreement statistics? Seems not found in the appendix?

Limitations:
Yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
Almost 50 pages of appendix.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes

Rebuttal by Authors

Rebuttal
by Authors (Jiajun Wu (/profile?id=~Jiajun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))
31 Jul 2025, 07:08 (modified: 31 Jul 2025, 14:42)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=xEqJq1K4Gq)

Rebuttal:
We thank Reviewer 5ZSD for their detailed and insightful feedback. We are particularly grateful that the reviewer found our code "very nicely structured," our illustrations "easy-to-understand," our benchmark framework "systematic," and our RL results showing "surprising effectiveness."

The primary concerns raised appear to stem from a focus on metric-level robotics applications. We have addressed all weaknesses and questions below.

Response to Weaknesses

1. **On the Toy-Map Abstraction.** We thank the reviewer for raising this concern. Below is why we focuse on the abstracted maps:
   - **Focus on Foundational Cognition, Not Metric Robotics**: Our primary goal is to test if VLMs can form coherent mental models of space, a cognitive task involving reasoning about unseen space across different views.
   - **Abstraction by Design to Isolate Variables**: We intentionally use a discrete grid to isolate this core spatial reasoning ability. This avoids conflating the results with errors from sensor noise or metric inaccuracies inherent in real-world SLAM pipelines.
   - **A Necessary Precursor to Complex Tasks**: We posit that robust abstract spatial understanding is a prerequisite for metric tasks. For instance, a system must first infer that an object exists behind a wall before it can attempt to localize it metrically [1].
   - **Enabling Rigorous & Focused Benchmarking**: This methodology allows for a clean and systematic evaluation of a VLM's ability to reason over an internal spatial model, providing a critical baseline for the field.
2. **On Idealised Visual Input.**
   - **We thank the reviewer for this important point on real-world robustness.** However, our primary aim is to isolate and benchmark the VLM's **core spatial reasoning ability**. Using controlled inputs ensures that observed failures stem from reasoning deficits, not from low-level perception errors.
   - **We also wish to clarify that our data, sourced from real-world datasets like ScanNet [2] and WildRGBD [3], does possess a degree of realism.** While view hops are aligned, these scenes contain authentic lighting, material textures, and other artifacts from the original 3D scans, providing a richer and more realistic environment than synthetic data.
   - **Ultimately, we aim to benchmark VLMs against a fair, human-centric standard.** Our goal is to test whether VLMs can form spatial models in scenarios where a human can. *Evaluating under extreme noise, where*

even humans would fail, would shift the focus from cognitive reasoning to perceptual resilience, which is a valuable but different research question.

3. **On the Static and Template-Driven QA.** We thank the reviewer for raising this critical point about overfitting.
   - **Our experiments were designed to test this exact hypothesis.** We found that forcing the model to first generate a map before answering (our "map-then-reason" paradigm) dramatically boosts accuracy from a 37.8% baseline to 76.1%.
   - **This major performance gain is inexplicable by linguistic overfitting alone.** Success in our task requires two key steps that go beyond pattern-matching: (1) synthesizing a single coherent map from multiple partial views, and (2) reasoning over the map's geometric structure to answer questions involving perspective-taking.
   - **Qualitative analysis provides further proof.** We observed that during the training, the cognitive map reconstruction quality boosts faster when combined with reasoning than training with the map itself (Figure 4 in our submission). This shows the map serves as a genuine reasoning scaffold, not just an intermediate text pattern.
   - **We agree that dynamic scenes and out-of-distribution language are important next steps.** Our work establishes a necessary foundation by proving that VLMs can form and effectively use a functional spatial representation in complex, static environments.
4. **On the Omission of State-of-the-Art Baselines.**
   - We have benchmarked the newer models as suggested. The new results have been added and underscore the benchmark's difficulty.

Model Name	Overall
o3-2025-04-16 (MindCube-Tinybench)	56.19
o4-mini-2025-04-16 (MindCube-Tinybench)	50.29

   - These models demonstrate improved performance. However, their overall accuracy still underscores the persistent challenges of our benchmark.
   - We have included these model results in our evaluation table.

Response to Questions

1. **What is the human performance on the benchmark?**
   - As noted in Appendix B.3, we conducted human performance on MindCube-tinybench. And the results are below:

Annotator ID	Overall Accuracy
GPT-4o	36.54
Human 1	92.82
Human 2	92.51
Human 3	92.36

   - Results demonstrate that average human performance is 92.5%, outperform any SOTA VLM by a large margin.
2. **What are the inter-annotator agreement statistics?**
   - **The inter-annotator agreement (IAA) for our QA annotations is over 98%**, which confirms the high quality and reliability of our benchmark data.
   - **We have added a section** to the appendix with a detailed breakdown of the annotation process and these IAA statistics.

[1] *Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age*. Cadena et al. IEEE Transactions on Robotics, 2016

[2] *ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes*. Dai et al. CVPR 2017

[3] *ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data*. Baruch et al. NeurIPS 2021

Replying to Rebuttal by Authors

Official Comment by Authors

Official Comment

by Authors (/profile?id=~Junjun_Wu1), Jieyu Zhang (/profile?id=~Jieyu_Zhang1), Manling Li (/profile?id=~Manling_Li1), Qineng Wang (/profile?id=~Qineng_Wang1), +10 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission4015/Authors))

03 Aug 2025, 12:09 (modified: 03 Aug 2025, 12:09) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewer 5ZSD,

Thank you again for your thoughtful and constructive review, and we would like to kindly follow up on your opinions on our rebuttal.

We worked extensively to address the points you raised, and we are keen to ensure our revisions are satisfactory and would be grateful for your feedback when you have a moment.

Thank you for your time and consideration.

The Authors