Spatial Mental Modeling from Limited Views

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans 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 beyonds 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 such scaffolding of spatial mental models, actively constructing and utilizing internal structured spatial representations with flexible reasoning processes, significantly improves understanding of unobservable space.

1 Introduction

For Vision-Language Models (VLMs) [1, 2, 3, 4] to move beyond passive perception [5, 6] to interact with partially observable environments [7, 8, 9], it is fundamental to reason about unseen spatial relationships from limited views. Consider how effortlessly a human can infer the unseen objects behind the "*plant*" are the "*tissue box*" and the "*hand sanitizer*" in the second viewpoint in Figure 1, including their position, pose, and their relationship with objects that are not simultaneously visible, all by integrating information from *four* ego-centric observations: we build and update a mental model of our surroundings, even when objects are out of sight. This is enabled by a core cognitive function known as the **spatial mental model** [10, 11]: an internal representation of the environment that allows for consistent understanding and inference about space, independent of the current viewpoint. VLMs, despite their impressive progress, struggle to synthesize spatial information from limited views, maintain spatial consistency across views, and reason about objects not directly visible [12, 13, 14, 15].

This gap calls for specialized evaluation settings, which must include: (a) reasoning with partial observations where objects are occluded or out of view (such as "*hand sanitizer*"in the second viewpoint in Figure 1), (b) maintaining cross-view consistency across shifting viewpoints (such as through anchor objects "*plant*"), and (c) mental simulation to infer hidden spatial relationships (such as "*what if turning left and moving forward*"). To fill this gap, we introduce MINDCUBE, featuring $17,530$ questions and $2,919$ images, organized into 740 multi-view groups through various types of

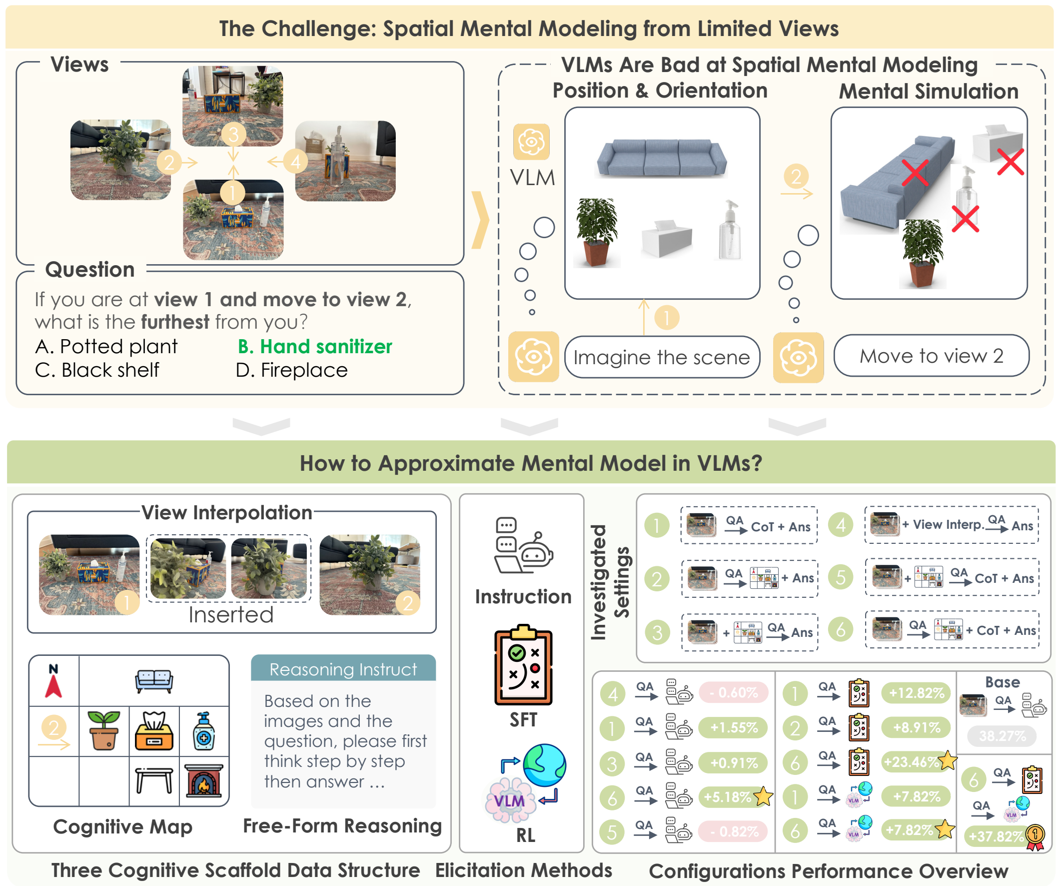


Figure 1: **Top**: VLMs cannot maintain a coherent mental model when evaluating on the MINDCUBE benchmark. **Bottom**: We study how we can help VLMs imagine space through external (scaling of views, cognitive map input) and internal strategies (fine-tuning, cognitive map elicitation). We find joint cognitive map and reasoning setting yields the highest gain (37.8%). ⭐: Best within the same elicitation method. 🏅: Best performance combination.

viewpoint transformations (i.e., ROTATION, AMONG, AROUND in Figure 2). We annotate questions with a focus on objects that are not visible in the current query view. As shown in Figure 2, we systematically design question types requiring "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*"), complex relation reasoning queries (referencing either the agent or other objects).

Our extensive evaluations of 14 state-of-the-art VLMs on MINDCUBE reveal that both open-source and closed-source models perform only marginally better than random guessing. This poor performance motivates a central question: **How can we help VLMs reason from partial observations**?

Inspired by spatial cognition [16, 17, 18] operating through *visual imagery*, *linguistic reasoning*, or *explicit cognitive maps*, to build consistent spatial awareness across different views, we investigate whether intermediate representations can help VLMs approximate mental models through three approaches. **View Interpolation** generates intermediate views between given observations using recorded video or 3D reconstruction (Stable Virtual Camera [19]), which unexpectedly is not helpful, highlighting the importance of reasoning directly from *limited* views. **Free-form Natural Language Reasoning** verbalizes the mental simulation process, achieving substantial gains (+1.6%). **Structured Cognitive Map** simulates global spatial memory from an allocentric (bird's-eye) perspective with orientation and view augmentation. Interestingly, ground truth cognitive maps yield minimal improvements (+0.9%), while guiding models to construct their own cognitive maps achieves better results (+5.2%), which indicates that teaching a model to *generate* its own mental model and think this way, is more effective than directly making sense of *provided* ready-made representations.

Despite generating seemingly correct maps, VLMs exhibit a significant bottleneck in *accurate* mental modeling, evidenced by low isomorphic similarity (17%) with ground truth maps. Recognizing

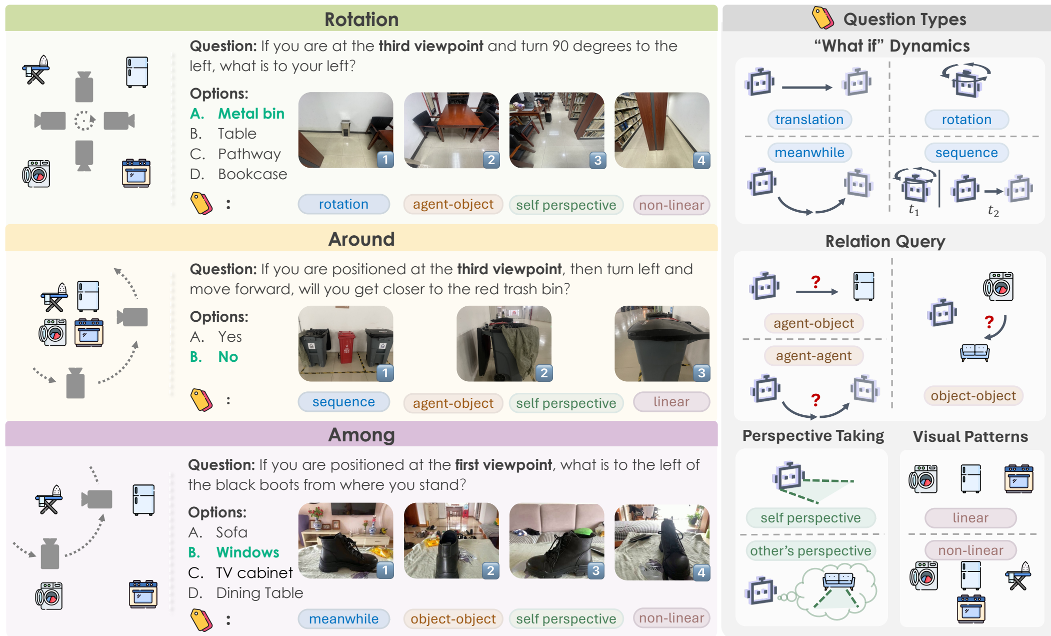


Figure 2: MINDCUBE taxonomy and examples. Left: Three camera movement patterns (ROTATION, AROUND, AMONG) with corresponding spatial QA examples. **Right**: Four-dimensional taxonomy categorizing MINDCUBE questions types.

this limitation, we train VLMs by constructing $10,000$ reasoning chains and $10,000$ ground truth cognitive maps, investigating how to effectively guide their thinking process through injection of these training signals. Self-supervised Finetuning (SFT) on cognitive maps significantly boost isomorphic similarity to 50% from 17%. While SFT on free-form reasoning chains proved more effective with a gain of $+4.8\%$, guiding models to first build cognitive maps and then perform free-form reasoning over them achieved significantly better performance, resulting in a total gain of $+15.4\%$, proving scaffolding spatial mental models via actively constructing and utilizing internal structured spatial representations with flexible reasoning processes is highly effective.

We employ Reinforcement Learning (RL) to further boost post-SFT performance, guiding models to think in terms of building and reasoning over cognitive maps by injecting structured thinking before RL training, using our SFT model. This approach leads to a significant improvement—raising task accuracy from a baseline of 38.3% to 76.1%. Our empirical evidence substantiates a critical finding: VLMs exhibit superior performance in spatial reasoning tasks when autonomously generating and leveraging internal mental representations, as compared to conventional approaches such as view interpolation or externally-supplied maps.

2 MINDCUBE Benchmark and Evaluation

2.1 MINDCUBE Benchmark

Overview. We introduce MINDCUBE, a benchmark for evaluating VLMs' spatial reasoning under partial observations and dynamic viewpoints. MINDCUBE features multi-view image groups paired with spatial reasoning questions, enabling fine-grained analysis of spatial modeling performance. It targets key challenges such as maintaining object consistency across views and reasoning about occluded or invisible elements. Table 1 (left) summarizes the benchmark's overall data distribution. Details on benchmark design, taxonomy, and curation are provided in the Appendix.

Taxonomy. For a fine-grained analysis of VLM spatial reasoning abilities, we introduce a taxonomy that systematically categorizes the challenges in MINDCUBE (visualized in Figure 2). This taxonomy spans five key dimensions: 1) **Camera Movement**: We mainly collect three types of camera movement: ROTATION (Stays in place but rotates to look around), AROUND (Moves around evaluated objects in a circular path), and AMONG (Moves among evaluated objects in a circular path). 2)

Table 1: Left: MINDCUBE data statistics. Right: Performance of VLMs on MINDCUBE.

Rotation(1081)	
Arkitscenes	865/53
Self collected	216/9
Img groups	62

Among(14782)	
WildRGB-D	13821/463
DL3DV-10K	961/30
Img groups	493

Around(1667)	
DL3DV-10K	725/109
Self collected	942/76
Img groups	185

Method	**Overall**	**Rotation**	**Among**	**Around**
Baseline				
Random (chance)	*39.40*	*36.36*	*39.00*	*46.44*
Random (frequency)	*41.60*	*39.00*	*39.00*	*46.00*
Open-source Multi Image Models				
LLaVA-Onevision-7B	**49.00**	36.84	**50.54**	38.43
LLaVA-Video-Qwen-7B	45.78	37.80	46.94	36.52
mPLUG-Owl3-7B	45.46	37.51	47.04	29.62
InternVL2.5-8B	29.01	35.69	27.00	52.97
Qwen2.5-VL-7B	35.44	38.09	34.73	45.13
LongVA-7B	35.29	35.31	34.95	40.55
idefics2-8B	38.74	37.22	38.65	41.83
DeepSeek-VL2-Small	47.42	36.36	49.44	28.03
Mantis-8B(Clip)	28.82	38.18	26.08	**61.36**
Proprietary Models(API)				
GPT-4o	36.70	**40.88**	36.07	42.89
Claude-3.7-Sonnet	39.71	37.70	39.62	43.42
Spatial Models				
RoboBrain	42.23	37.51	42.75	39.28
space-mantis	31.56	37.99	29.41	58.07
space-LLaVA	27.31	33.89	26.71	35.37

Visual Patterns: This describes the objects' spatial configurations, including spatial *linear* or *non-linear* arrangements. 3) **"What-if" Dynamics**: The hypothetical transformations applied to the agent's viewpoint, such as *translation*, *rotation*, or their combination (*meanwhile* and *sequence*). 4) **Relation Query**: The type of spatial relation being queried, including *agent–object*, *agent–agent*, or *object–object*. 5) **Perspective Taking**: Whether the spatial reasoning is grounded in the perceiver's own viewpoint (*self*) or involves adopting the viewpoint of another entity (*other*).

Dataset Curation. The MINDCUBE dataset was created through a pipeline: We first selected multi-view image groups matching our taxonomy's movement patterns (Figure 2) and spatial criteria. These were then annotated with key spatial information. Finally, we algorithmically generated taxonomy-aligned questions with targeted distractors. Details are included in the Appendix.

2.2 Evaluation on MINDCUBE

We evaluate VLMs' spatial reasoning on MINDCUBE using a diverse set of models (Table 1, right; setup details in the Appendix). Results reveal a striking performance gap: the best model, LLaVA-Onev-7B, achieves only 49.00% accuracy—well above chance but far from human-level. ROTATION tasks proved hardest (top score: 40.88%), suggesting limited mental rotation and viewpoint adaptation. AMONG and AROUND tasks showed no consistent winners, highlighting weak relational reasoning. Large proprietary models often lagged behind smaller open-source counterparts. Spatial fine-tuning yielded mixed results. Overall, neither multi-image input nor spatial fine-tuning reliably improves spatial reasoning, raising a key question: **How can we help VLMs develop or approximate these crucial spatial reasoning capabilities?**

3 Which Scaffolds Best Guide Spatial Thinking in Unchanged VLMs?

To address the identified gap, we first evaluate whether structured data forms can scaffold spatial reasoning in frozen VLMs by approximating spatial mental models under limited views.

3.1 Data Structures as Cognitive Scaffolds for Spatial Mental Models

We investigate whether certain data structures can act as cognitive scaffolds that help VLMs form spatial mental models from limited visual observations. In cognitive science, spatial mental models are internal representations encoding the relative configuration of objects and viewpoints. Rather than metric-precise maps, they are schematic, manipulable constructs that support reasoning across fragmented observations and unseen perspectives [11, 20, 21, 22]. For instance, humans can mentally simulate turning or infer what lies behind them, suggesting that such representations are flexible, incomplete, yet functionally effective. Drawing on this literature, we define three data structures,

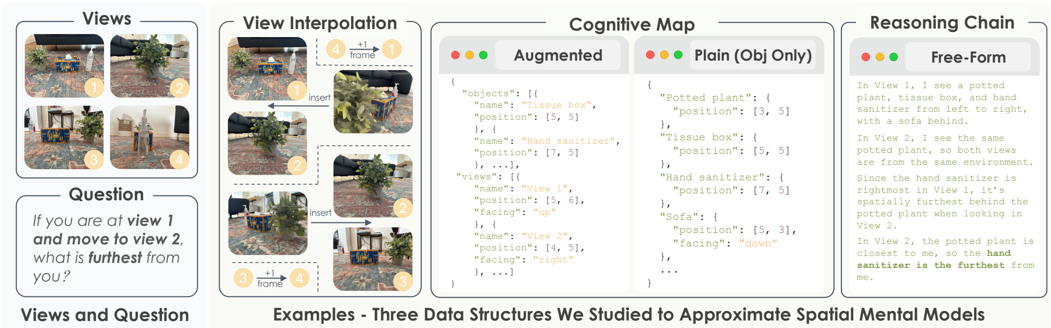


Figure 3: Grounded examples of our three data structures that approximate spatial mental models.

each targeting distinct cognitive properties (integration, transformation, inference) of spatial mental models, with grounded examples in Figure 3:

1. **View Interpolation**. Interpolating between sparse camera views introduces perceptual continuity, echoing the process of *mental animation* [23] and supporting internal transformation such as imagined rotation. This structure scaffolds the dynamic updating capability of spatial mental models. Figure 3 shows a one-frame inserting example that replaces the original question images.
2. **Augmented Cognitive Map**. A cognitive map is a 2D schematic representation of object layouts in space. Such maps resemble Tversky's *cognitive collages* [20], and they capture locally coherent but fragmented structures. Recent studies [24, 25] on VLM-based spatial intelligence typically adopt a *plain* form that only encodes object positions in a top-down view. We propose an *augmented* variant that incorporates discrete views, with both objects and views annotated by position and orientation, thereby approaching the relational consistency of *spatial mental models*.
3. **Free Form Reasoning**. Open-ended, step-by-step natural language reasoning offers a *procedural approximation* of how spatial models are constructed and queried. While less rigid than map-like structures, such reasoning reflects the inferential function of spatial mental models, especially under ambiguous or incomplete observations [21].

3.2 Experiment Setup

We conduct controlled experiments with fixed input formats to test whether structured scaffolds can help without retraining. Each condition introduces a different structure to support internal modeling under limited views.

Table 2: Abbreviations for the ten input-output configurations across all experiments in this work. VI = **V**iew **I**nterpolation, CGMap = **C**o**g**nitive **Map**, Aug = Augmented (objects + camera included), FF-Rsn and FFR = Free-Form Reasoning.

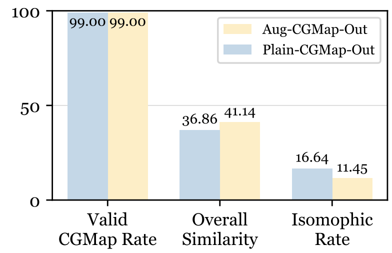
Name	Input Structure	Output Format
`Raw QA`	Raw views + question	Direct answer
`VI-1`	Raw + 1 interp. view	Direct answer
`VI-2`	Raw + 2 interp. views	Direct answer
`FF-Rsn`	Raw views + question	Reasoning → answer
`Aug-CGMap-In`	Aug. cognitive map	Direct answer
`Aug-CGMap-Out`	Aug. cognitive map	Direct answer
`Plain-CGMap-Out`	Plain cognitive map	Direct answer
`Plain-CGMap-FFR-Out`	Raw views + question	Plain map + rsn → ans
`Aug-CGMap-FFR-Out`	Raw views + question	Aug. map + rsn → ans
`CGMap-In-FFR-Out`	Aug. cognitive map	Reasoning → answer

Model and Evaluation Data We conduct all experiments using *Qwen2.5-VL-3B-Instruct* [3]. Our evaluation is performed on MindCube-Tiny, a diagnostic subset sampled from MindCube, containing 1,100 questions in total. Detailed statistics are: 500 from the Among, 400 from Around, and 200 from Rotation.

Configurations Each experiment is defined by two orthogonal axes: *Input Structure* (what spatial evidence VLMs receive) and *Output Format* (the required response type). We investigate a subset of the ten possible configurations shown in Table 2. Specifically, our grounded cognitive maps are generated using the object arrangements annotation described in Section 2.1, and examples for all configurations are provided in the Appendix. We exclude the `Aug-CGMap-Out` and `Plain-CGMap-Out` settings, as VLMs tend to conflate map generation with reasoning, even when instructed otherwise.

Table 3: Left: QA accuracy (%) of *Qwen2.5-VL-3B-Instruct* on the MINDCUBE-TINY benchmark under different input/output scaffolds. Right: Graph metrics for two cog map output settings.

Config.	Overall	Rotation	Among	Around
`Raw QA`	38.27	35.50	31.40	48.25
`VI-1`	37.67↓	30.05	32.80	47.25
`VI-2`	36.10↓	22.95	31.80	47.50
`Aug-CGMap-In`	39.18↑	38.00	31.60	**49.25**
`FF-Rsn`	39.82↑	33.50	38.40	44.75
`Aug-CGMap-FFR-Out`	42.73↑	37.50	**44.60**	43.00
`Plain-CGMap-FFR-Out`	**43.45**↑	**42.50**	41.00	47.00
`CGMap-In-FFR-Out`	37.45↓	35.00	31.60	46.00



Evaluation Metrics We evaluate task performance using QA accuracy. For generated cognitive maps, we introduce a set of well-defined graph metrics: (1) *Valid Cognitive Map Rate*, indicating whether the output conforms to the expected schema; (2) *Overall Similarity*, a weighted score combining directional and facing consistency; and (3) *Isomorphism Rate*, measuring whether all pairwise object relations match the ground truth under optimal alignment. Full definitions are provided in Appendix.

3.3 Do Scaffolds Improve Spatial Reasoning Without Training?

We evaluate how well the seven input configurations defined in Table 2 support spatial reasoning in VLMs under limited views, without any model updates. Results are shown in Table 3 (left).

How far can structure alone go? We begin with the baseline: raw input views and direct answering, which achieves just 38.27% accuracy. Adding interpolated views, which we hope to simulate smoother perceptual transitions, leads to no meaningful gain, and in some cases slightly harms performance (down to 36.10%). Providing an augmented cognitive map or free-form reasoning barely improve accuracy (39.18% and 39.82%). These results suggest: *structure alone, whether visual or spatial, is not enough*. Without engaging reasoning, VLMs struggle to leverage even well-formed spatial cues.

Can we prompt the model to think spatially? The answer appears to be yes. Strikingly, compared to `FF-Rsn` only, prompting models to produce a *cognitive map* before answering leads to the strongest gains: 43.45% with plain maps (object only), and 42.73% with augmented maps (objects + views). This suggests that generating a map may encourage the model to first form a global understanding of the scene, which in turn supports more structured reasoning. Our case studies in the Appendix reveal that such reasoning often unfolds on the map itself. Both map forms have a great format following ability. Augmented maps perform slightly worse. In Table 3 (Right), despite generating syntactically valid maps for both formats, similarity to grounded maps is low ($< 50\%$), reflecting limited mapping ability. Notably, augmented maps have lower isomorphism rates (↓ 5.19%), likely because the added view-level details increase generation errors, making downstream reasoning less reliable.

What if we ask models to reason over grounded cognitive maps? Intuitively, combining map input and free-form reasoning seems promising. Yet, performance drops to 37.45%. We hypothesize this mismatch arises from representational dissonance: reasoning over externally imposed maps may conflict with the model's latent space, leading to confusion rather than clarity.

Key Takeaways: Scaffolding Spatial Reasoning in Frozen VLMs

- *Producing reasoning is effective, especially combined with cognitive maps.*
- *Self-generated reasoning consistently outperforms imposed formats.*
- *Visual continuity and passive structure provide little benefit.*
- *Simpler, object-only maps work better; too much structure can backfire for frozen VLMs.*

4 Can We Teach VLMs to Build and Leverage Spatial Representations?

So far, prompting frozen VLMs with external scaffolds, such as interpolated views or cognitive maps, has yielded limited gains. These techniques fail to tackle the core limitation: VLMs do not form internal spatial representations or reason through space effectively. To go further, we want to know: Can supervised fine-tuning (SFT) teach VLMs to build and leverage spatial models from within?

Table 4: QA accuracy (%) and cognitive map generation quality of *Qwen2.5-VL-3B-Instruct* under SFT configurations on MINDCUBE-TINY. Both `FF-Rsn` and `FFR` refer to free-form reasoning.

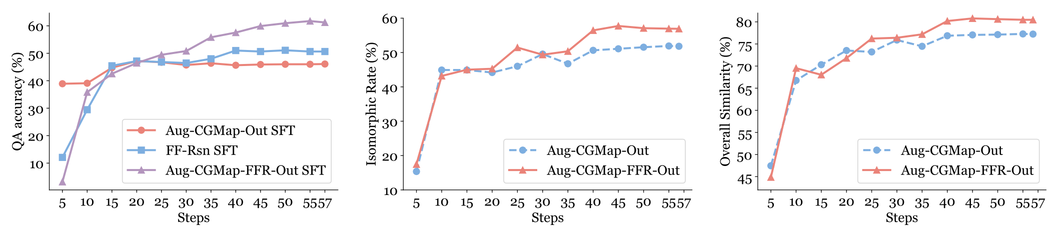
SFT Config.	MINDCUBE-TINY QA Accuracy (%)				Generated Cognitive Map (%)		
	Overall	**Rotation**	**Among**	**Around**	**Valid Rate**	**Overall Sim.**	**Isom. Rate**
`Raw QA`	46.36	33.50	51.20	46.75	–	–	–
`Aug-CGMap-Out`	47.18$\uparrow$	35.00	52.80	46.25	100.00	77.27	51.91
`Plain-CGMap-Out`	45.73$\downarrow$	37.50	49.80	44.75	100.00	75.89	49.55
`FF-Rsn`	51.09$\uparrow$	34.00	56.40	**53.00**	–	–	–
`Aug-CGMap-FFR-Out`	**61.73**$\uparrow$	**68.50**	**70.60**	47.25	**100.00**	**80.46**	**56.91**

4.1 Designing a Robust Experimental Framework

To ensure consistency and comparability, we inherit experimental configurations detailed in Sections 3.1 and 3.2. Specifically, we retain: (1) the two effective data structures—Cognitive Maps (Object-only / Object + Camera) and Free-Form Reasoning, (2) the base model *Qwen2.5-VL-3B-Instruct*, (3) the evaluation benchmark MINDCUBE-TINY, and (4) all established evaluation metrics. View interpolation is excluded from our fine-tuning experiments due to its limited performance gains in earlier validations. Primary modifications in this SFT phase include adjusted training hyperparameters (detailed in the Appendix).

SFT Task Configurations Drawing on insights from Section 3.3, we use selected configurations from Table 2 to evaluate the incremental impact of cognitive map generation and free-form reasoning in SFT. These include baseline QA without explicit reasoning (`Raw QA`), reasoning guided by generated maps only (`Plain-CGMap-Out`, `Aug-CGMap-Out`), reasoning-augmented prompts (`FF-Rsn`), and a fully integrated setup that asks VLMs to generate both maps and reasoning (`Aug-CGMap-FFR-Out`).

Grounded Free-Form Reasoning Chain Generation We design grounded reasoning chains using detailed image annotations and structured question templates. Chains are manually constructed via a template-based method, ensuring logical coherence and clear grounding in observable spatial relations (see an example in Figure 3). This yields precise, interpretable supervision signals that help VLMs learn robust spatial reasoning representations. The detailed grounded reasoning data generation pipeline is shown in the Appendix.



(a) QA accuracy (%) (b) Isomorphic Rate (%) (c) Overall Similarity (%)

Figure 4: SFT per 5 step training performance on task accuracy and graph metrics.

4.2 Do VLMs Truly Benefit from Explicit Training in Spatial Reasoning?

We explore several SFT configurations (results shown in Table 4), guided by a series of core questions. Fine-tuning directly on raw QA pairs, without spatial supervision, raises accuracy from 38.27% to 46.36%. This suggests VLMs can absorb some spatial cues from QA data alone. We use this setup as the baseline for evaluating methods that explicitly incorporate spatial structures.

Can structured approximations of mental models alone meaningfully improve performance? As shown in Table 2, supervised fine-tuning on explicit cognitive maps, either *Augmented* or *Plain*, leads to substantial improvements in graph structure quality, with more than 30% gains in both overall similarity and Isomorphic rate. However, the effect on end-task accuracy remains limited. Both augmented maps (47.18%) and Plain maps (45.73%) offer only modest gains over the fine-tuned `Raw QA` (46.36%). In contrast, directly `FF-Rsn` yields a significantly larger improvement, increasing accuracy to 51.09%. This could be that free-form reasoning is well-aligned with base model language ability, which better guides the model towards the correct spatial understanding.

Generating both cognitive maps and free-form reasoning is the most effective approximation. Among all configurations, this combination yields the most significant performance gain (61.73%,

↑15.37%), far surpassing models that rely on either map generation or reasoning alone. This suggests **a strong synergy** between structured spatial modeling and natural language inference. Why does this combination work so well? First, improvements in task accuracy are accompanied by sharper spatial representations. The `Aug-CGMap-FFR-Out` model, which first builds a cognitive map and then reasons over it, achieves 80.46% similarity and 56.91% isomorphism, surpassing all other variants. On the other hand, training dynamics further reinforce this view. As shown in Figure 4, models trained to jointly map and reason tend to converge more slowly, but ultimately achieve higher performance. Interestingly, spatial representation quality improves more quickly when reasoning is explicitly trained. These results suggest that reasoning encourages more precise and structured spatial understanding, rather than merely consuming it.

Key Takeaways: Teaching VLMs to Reason Spatially

- *Joint cogmap and reasoning setting yields optimal performance through synergistic effects.*
- *Reasoning improves accuracy and map quality by strengthening spatial representations.*
- *More structured outputs slow convergence but lead to higher final performance.*

5 Can Reinforcement Learning Further Refine Spatial Thought Processes?

While SFT establishes a strong baseline for spatial reasoning, emerging evidence from models like DeepSeek R1 [26, 27] suggests reinforcement learning (RL) can offer additional gains by optimizing behavior through outcome-driven feedback. We ask: Can reward-guided refinement help VLMs build sharper spatial models and reason more effectively?

5.1 Experimental Setup

We employ the VAGEN framework [28] for VLM policy optimization, using Group Relative Policy Optimization (GRPO)[29] as our core algorithm. To manage compute cost, we train each configuration for only 0.5 epoch. For fair comparison, the RL setup retains all key components from the SFT stage, including the base model, spatial input formats, benchmark dataset (MINDCUBE-TINY), and evaluation metrics, as detailed in Sections 3.1 and 3.2. Additional details appear in the Appendix.

Task Configurations and Reward Design. We evaluate three RL variants: (1) `RL-FF-Rsn (from scratch)`, which trains *Qwen2.5-VL-3B-Instruct* to produce free-form reasoning chains; (2) `RL-Aug-CGMap-FFR-Out (from scratch)`, which trains the model to jointly generate cognitive maps and reasoning; and (3) `RL-Aug-CGMap-FFR-Out (from SFT)`, which initializes from the strongest SFT checkpoint. The reward function is sparse but targeted: +1 for structurally valid outputs, and +5 for correct answers.

Table 5: QA accuracy (%) and cognitive map generation quality of *Qwen2.5-VL-3B-Instruct* under various RL configurations on MINDCUBE-TINY.

RL Config.	MINDCUBE-TINY QA Accuracy (%)				Generated Cognitive Map (%)		
	Overall	Rotation	Among	Around	Valid Rate	Overall Sim.	Isom. Rate
`RL-FF-Rsn (from scratch)`	46.09	32.50	53.40	43.75	–	–	–
`RL-Aug-CGMap-FFR-Out (from scratch)`	47.91	35.00	53.40	47.50	99.73	65.73	0.00
`RL-Aug-CGMap-FFR-Out (from SFT)`	**76.09**	**78.50**	**96.80**	**49.00**	**100.00**	**84.51**	**66.36**

5.2 Can Reinforcement Learning Unleash the Power of Approximating Spatial Mentaling?

Reinforcement learning (RL) lets a model *feel* the consequences of its spatial thoughts through reward, but does that feedback alone forge a genuine "mental map," or must we first teach the model what a map looks like? Table 5 summarizes three key settings and answers the question in two parts.

RL in a vacuum is not enough. Training `RL-FF-Rsn` from scratch and asking the policy to emit free-form reasoning and rewarding only correct answers barely surpasses the raw QA baseline (46.09% overall; 32.50% on the hardest *Rotation* queries). Sparse rewards guide the model toward useful heuristics, yet provide too little structure for constructing a robust spatial abstraction.

Structured outputs provide modest benefits when learned from scratch. Adding cognitive-map generation (`RL-Aug-CGMap-FFR-Out`) introduces an explicit spatial scaffold the policy must satisfy. Accuracy nudges to 47.91%, and map validity soars to 99.73%. Still, without a prior notion of "good" geometry, RL cannot fully exploit the scaffold: similarity and isom. remain low (65.73%, 0.00%).

RL shines when it stands on an SFT-built scaffold. Warm-starting RL from the optimal SFT checkpoint (`RL-Aug-CGMap-FFR-Out (from SFT)`) produces remarkable improvements: 76.09% overall QA accuracy, representing a $\uparrow$ **14.36%** gain over SFT alone and $\uparrow$ **29.00%** over RL-from-scratch approaches. Spatial metrics show parallel enhancements—map similarity reaches 84.51% while isomorphism hits 66.36%, nearing oracle-level performance. Qualitative analysis reveals the policy effectively eliminates extraneous objects, generates streamlined maps, and produces concise yet definitive reasoning chains. These results suggest RL is (1) *polishing* rather than reconstructing spatial representations from the ground up, and (2) *raise the convergence ceiling* of SFT, enabling the model to break through previous performance plateaus.

Key Takeaways: Reinforcement Learning for Spatial Reasoning

- *RL enhances spatial reasoning, even without prior supervised fine-tuning.*
- *Combining cognitive maps with reasoning consistently improves all learning outcomes.*
- *RL best boosts performance post-SFT, pushing the upper limits of spatial reasoning.*

6 Related Works

Spatial Cognition. Spatial cognition encompasses skills like mental rotation, spatial visualization and object assembly, essential for perceiving and manipulating spatial relationships in both 2D and 3D environments [30, 18, 31]. At the core of these abilities are Spatial Mental Models (SMMs) [10, 11], which are internal representations that allow for consistent understanding about space. Recently, much effort has been dedicated to evaluating spatial cognition in VLMs [32, 12, 17, 33]. Moreover, some methods are proposed to enhance spatial understanding such as coordinate-aware prompting [34], CoT reasoning [9, 35], explicit spatial representation alignment [36, 37], and RL-based approach [38]. However, existing benchmarks and approaches often neglect the mental-level spatial reasoning that underpins human cognition, leaving a gap between machine and human capabilities. To bridge this gap, a new approach is needed that trains VLMs to reason about space not only through visual data but also through mental-level spatial reasoning, aligning more closely with human spatial cognition.

Multi Views understanding. Prior work in multi-view fusion has explored 3D perception by reconstructing point clouds from images captured along random trajectories [39, 40, 41]. These methods focus on creating 3D models but often lack a detailed understanding of object-specific spatial relationships, particularly in terms of maintaining spatial consistency across multiple views. On the other hand, existing multi-image or video benchmarks [42, 43, 44, 45, 46] primarily focus on assessing perceptual and reasoning abilities, falling short in thoroughly evaluating spatial consistency and the ability to maintain coherent scene representations. Recently, some spatial multi-view benchmarks [47, 15, 14, 8, 13, 48, 7] aim to assess spatial reasoning in multi-view settings. However, they often overlook the fusion and consistent representation of spatial information across different perspectives. To address this gap, we propose MINDCUBE to bridge the divide between perception and spatial consistency understanding.

7 Conclusion and Future Impact

We introduced MINDCUBE to study how VLMs can approximate spatial mental models from limited views, a core cognitive ability for reasoning in partially observable environments. Moving beyond benchmarking, we explored *how* internal representations can be scaffolded through structured data and reasoning. Our key finding is that *constructing and reasoning over self-generated cognitive maps*, rather than relying on view interpolation or externally provided maps, yields the most effective approximation of spatial mental models across all elicitation methods (input-output configurations, supervised fine-tuning, and reinforcement learning). Initializing RL from a well-trained SFT checkpoint further optimizes the process, pushing spatial reasoning performance to new limits.

Future Impact. Our work establishes that combining cognitive map generation with reasoning to model spatial information is the most effective. We believe that once high-quality SFT datasets for cogmap generation and reasoning are established, RL can be leveraged to further push the performance boundaries. We anticipate the exploration of novel training paradigms designed to unlock even greater synergistic effects and thus achieving a "1+1 > 2" impact on spatial intelligence.

References

[1] OpenAI. Hello gpt-4o. Blog, 05 2024. Accessed: November 22, 2024. 1

[2] Anthropic. Claude 3.5 sonnet. Blog, 10 2024. Accessed: November 22, 2024. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 5

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 1

[7] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2024. 1, 9

[8] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities, 2025. 1, 9

[9] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jieneng Chen, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning, 2025. 1, 9

[10] Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980. 1, 9

[11] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983. 1, 4, 9

[12] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark, 2025. 1, 9

[13] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d, 2025. 1, 9

[14] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models, 2025. 1, 9

[15] Yiqi Zhu, Ziyue Wang, Can Zhang, Peng Li, and Yang Liu. Cospace: Benchmarking continuous space perception ability for vision-language models, 2025. 1, 9

[16] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models?, 2025. 2

[17] Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation, 2025. 2, 9

[18] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm, 2025. 2, 9

[19] Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 2

[20] Barbara Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *European conference on spatial information theory*, pages 14–24. Springer, 1993. 4, 5

[21] Barbara Tversky, Nancy Franklin, Holly A Taylor, and David J Bryant. Spatial mental models from descriptions. *Journal of the American society for information science*, 45(9):656–668, 1994. 4, 5

[22] Barbara Tversky. Structures of mental spaces: How people think about space. *Environment and behavior*, 35(1):66–80, 2003. 4

[23] Mary Hegarty. Mental animation: Inferring motion from static displays of mechanical systems. *Journal of experimental psychology: learning, memory, and cognition*, 18(5):1084, 1992. 5

[24] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 5

[25] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Rouyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 5

[26] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8

[27] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025. 8

[28] Kangrui Wang*, Pingyue Zhang*, Zihan Wang*, Qineng Wang*, Linjie Li*, Zhengyuan Yang, Chi Wan, Yiping Lu, and Manling Li. Vagen: Training vlm agents with multi-turn reinforcement learning. `https://mll-lab.notion.site/vagen`, 2025. Notion Blog. 8

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 8

[30] Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics, 2025. 9

[31] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation, 2025. 9

[32] Weichen Zhan, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space, 2025. 9

[33] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation, 2025. 9

[34] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 9

[35] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, Helong Huang, Guangjian Tian, Weichao Qiu, Xingyue Quan, Jianye Hao, and Yuzheng Zhuang. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning, 2025. 9

[36] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024. 9

[37] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. 9

[38] Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025. 9

[39] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images, 2023. 9

[40] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. 9

[41] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019. 9

[42] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, 2024. 9

[43] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024. 9

[44] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context, 2024. 9

[45] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024. 9

[46] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues, 2025. 9

[47] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms, 2025. 9

[48] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. 9

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope by identifying a significant gap in VLMs, introducing a new benchmark, evaluating three approaches, and highlighting a key insight that significantly improves understanding of unobservable space, all of which are well-supported by the paper's content and findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of the work about spatial mental modeling in Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We detail the assumption and proof of theoretical result on time complexity in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We presented comprehensive experimental information in paper, ensuring that our results are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we plan to open-source our data and code to promote the development of the field of spatial intelligence. We have also provided them in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All this information is presented in detail in our supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: LLM training and inference tasks are very resource intensive and costly.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we conducted our experiments primarily on two H100 GPUs, and you can find more details in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[Yes]

    Justification: You can find more details in the supplementary materials about the discuss of our self collected data.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not foresee any high risk for misuse of this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we credited them in appropriate ways.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we provided it in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In the experimental section, we described the large models we used to test performance.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.