Reasoning Path and Latent State Analysis for Multi-view Visual Spatial Reasoning: A Cognitive Science Perspective

Qiyao Xue[1*], Haoming Wang[1*], Weichen Liu[1], Shiqi Wang[1], Yuyang Wu[2], and Wei Gao[1]

[1] University of Pittsburgh, Pittsburgh, PA, USA
{qiyao_xue, hw.wang, weichen.liu, shw322, weigao}@pitt.edu
[2] Carnegie Mellon University, Pittsburgh, PA, USA
yuyangwu@andrew.cmu.edu

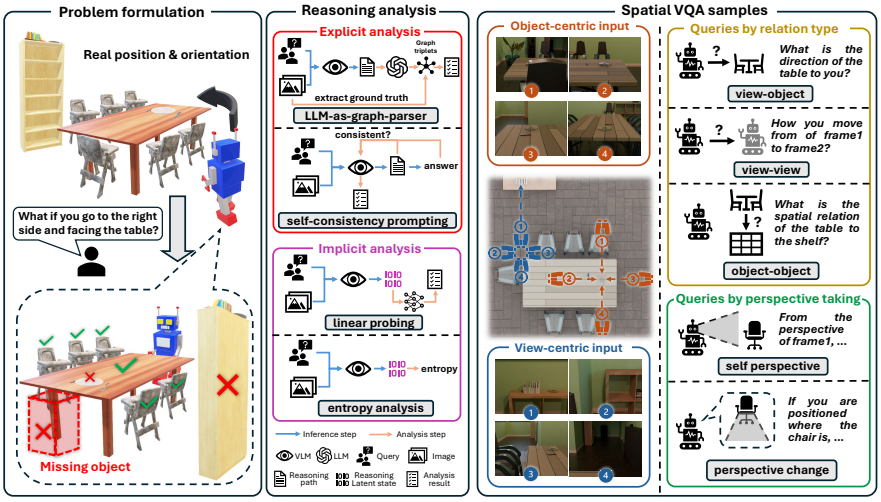

Fig. 1: Left: Current VLMs struggle to maintain coherent spatial reasoning across multiple views (✓ indicates consistent and × denotes incorrect reasoning). **Middle:** To assess VLM's successes and failures, we conduct explicit analysis of VLM's textual reasoning path and implicit analysis of VLM's latent token representations. **Right:** Such analysis builds on properly designed VQA samples that cover diverse viewpoint spatial patterns and query types to capture key cognitive factors in spatial reasoning.

Abstract. Spatial reasoning is a core aspect of human intelligence for perception and inference in 3D environments. However, current vision-language models (VLMs) struggle to maintain geometric coherence and cross-view consistency for spatial reasoning in multi-view settings. In this paper, we aim to understand the successes and failure of VLMs in multi-view spatial reasoning, via structured analysis of how VLMs encode perceptual evidence, integrate relational information, construct in-

* Equal contribution.

termediate spatial representations and perform perspective transformation. Guided by cognitive science theories of human spatial reasoning, our analysis examines VLM's consistency and representational persistence across reasoning phases that align with human's spatial cognition, in explicit and implicit manners. Using a self-constructed spatial benchmark, namely *ReMindView-Bench*, which consists of >50,000 VQA samples with complementary views and controlled viewpoint configurations, spatial patterns and query types, our explicit analysis using an LLM-as-graph-parser workflow and self-consistency prompting shows that VLMs remain reliable for in-frame perceptual encoding but degrade sharply when integrating information across views. Our implicit analysis of linear probing and entropy dynamics further reveals progressive loss of task-relevant spatial information and increasing uncertainty over reasoning phases. These findings provide important insights of how multi-view spatial mental models are formed, destabilized, and degraded across reasoning phases in VLMs. The ReMindView-Bench benchmark is available at `https://huggingface.co/datasets/Xue0823/ReMindView-Bench`, and the source codes of benchmark construction and VLM reasoning analysis are available at `https://github.com/pittisl/ReMindView-Bench`.

1 Introduction

Spatial reasoning is essential to infer spatial relations, geometric configurations, and real-world object dynamics that are not directly observable [12, 47], thereby supporting higher-level cognitive functions including navigation, manipulation, and world-model construction [25, 31]. In particular, spatial reasoning in multi-view visual settings, where reasoning needs information from multiple viewpoints, is more difficult, because it requires integrating observations from different viewpoints while maintaining geometric coherence and cross-view consistency [15, 59, 62]. To meet these requirements, models must reconstruct scene layouts and infer relative positions, orientations, and occlusions across perspectives [11, 15, 56, 61], and this demands stable mental model construction and perspective transformation rather than isolated perceptual recognition.

Although recent advances in Vision-Language Models (VLMs) have improved multimodal perception and compositional reasoning, they mostly rely on superficial cues such as color patterns, background regularities, or 2D co-occurrence statistics when inferring spatial relations [28,59,65]. Hence, as shown in Figure 1 - left, they often produce inconsistent reasoning across views, including missing objects and incorrect cross-view alignment. These behaviors reveal a persistent gap between human-like spatial cognition and current VLM's abilities [17,21,45].

To understand the successes and failures of VLMs in multi-view spatial reasoning, measuring task accuracy is insufficient. Instead, in this paper we focus on structured analysis of how VLMs encode perceptual evidence, integrate relational information, construct intermediate spatial representations, and perform perspective transformation. Being different from existing analysis that evaluates spatial reasoning as *a single monolithic pass* [43,48,66], our analysis is guided by cognitive science theories of human-like spatial reasoning [16,33,49] that involves

sequential processes of perceptual encoding, relational integration, mental model construction, and mental transformation [13, 18, 19, 24, 42, 46], and correspondingly examines the VLM's consistency and representational persistence across these reasoning phases that align with human's spatial cognition.

As shown in Figure 1 - middle, our analysis includes both explicit and implicit manners. Explicitly, we assess the correctness of VLM's reasoning path over different phases by using LLMs to parse the VLM's rich metadata of scene graphs generated in reasoning [43, 48], inspired by Johnson-Laird's theory of mental models [18, 19]. We further apply self-consistency prompting to evaluate whether VLM's generated rationales align with its internal decision-making [29, 50, 51]. Implicitly, we probe token-level latent representations using linear probing [32, 44] to determine whether task-relevant information persists across reasoning phases, and analyze entropy dynamics [22, 55, 58] to quantify uncertainty and calibration as reasoning progresses.

The effectiveness of such analysis builds on properly designed VQA samples, which cannot be found in the existing spatial datasets and benchmarks. Single-view datasets provide only partial context and miss occlusions and hidden relations across perspectives [5, 7, 17, 35, 65]. Video-based datasets contain multiple frames, but lack structured viewpoint control to examine VLM's abilities in aligning and maintaining spatial mental models across complementary viewpoints [8, 28, 37, 56]. Multi-view datasets [15, 57, 59, 64, 66] treat different views independently and hence cannot be used to examine cross-view coherence in spatial reasoning. They also sample viewpoints from fixed trajectories or real-world captures, resulting in uncontrolled variability in camera geometry and limited disentanglement of cognitive spatial factors. More discussions about those existing works can be found in Appendix J.

To address these limitations, as shown in Figure 1 - right, we constructed a new spatial reasoning benchmark, namely RemindView-Bench, which contains VQA samples with multiple complementary views [8,28,37,56] along various controlled dimensions, including different spatial patterns [3,46] (e.g., object-centric vs. view-centric) and object relationship being queried [13, 14, 24, 42] (e.g., self-perspective versus perspective-changing, object-object, view-object and view-view). Placement and orientation of camera, object distances, occlusion patterns and viewpoint geometry are explicitly parameterized and controlled. Hence, we isolate sources of spatial difficulty and enable systematic manipulation of cross-view alignment, perspective transformation, and relational integration factors [16, 33, 49], attributing VLM's reasoning errors to specific spatial factors.

Using >50,000 VQA samples constructed as described above, we conducted comprehensive analysis over 15 VLMs. Results show that current VLMs perform far below the human level, with task accuracies plateauing at ∼30-45% compared to 81.5% for humans. Performance systematically favors object-centric over view-centric conditions, degrades under counterfactual and cross-frame queries, and declines with increasing scene complexity. Phase-wise reasoning analysis further shows that models remain relatively stable during early perceptual encoding but deteriorate during relational integration and perspective transformation. Entropy trajectories indicate rising uncertainty and poor calibration in later rea-

soning stages. These findings provide new insights into the internal dynamics of VLM spatial reasoning. Rather than failing solely due to task difficulty, VLMs exhibit phase-specific breakdowns in geometric alignment and cross-view coherence. By linking explicit reasoning structures with implicit latent behaviors, our analysis enables attribution of reasoning errors to distinct cognitive processes, and this cognitively grounded diagnosis offers principled guidance for improving multi-view spatial reasoning and world-model consistency in future VLMs.

2 Preliminaries

2.1 Spatial Reasoning in VLMs

Spatial reasoning remains a core challenge for VLMs, as it requires grounding linguistic expressions into geometric and relational structures of physical world rather than symbolic sequences. This explains why VLMs are incapable of genuine spatial understanding but rely on language priors or dataset bias [5, 15, 59].

Achieving robust spatial reasoning would require human-like spatial perception and mental simulation [18, 42]. Although spatial reasoning of VLMs has been enhanced by integrating explicit geometric priors [5, 67], multi-view consistency learning [15, 54] and spatial reasoning modules [7, 34] to capture spatial understanding beyond single-image semantics, current VLMs still lack 3D internal representations for flexible viewpoint transfer [56, 57]. Cognitively grounded benchmarking and analysis are needed to align spatial reasoning tasks with mechanisms of human perception, construction, and mental transformation.

2.2 The Cognitive Perspective in Spatial Reasoning

Our construction of VQA samples and VLM reasoning analysis both build on the cognitive perspective of humans' spatial understanding.

Construction of VQA samples. Research in visual cognition suggests to distinguish between *object-centric* and *view-centric* representations, which structure spatial attention and memory in different ways. Object-centric representation integrates multiple views of the same object, maintaining geometric identity across views [20,41]. View-centric representation reflects spatial organization tied to the observer's position and viewing direction, producing context-dependent scene content [3, 46]. This dichotomy motivates our dual visual design in Figure 1, where object-centric inputs test geometric integration across views and view-centric inputs evaluate perceptual stability under view changes.

We incorporate the scene variability in layout diversity and object densities to reflect the cognitive principles in human spatial reasoning. According to the schema-based effect on spatial memory [49], humans encode and recall spatial layouts based on high-level scene schemas (e.g., room type or functional zones) rather than absolute coordinates, and we hence incorporate diverse types of indoor rooms to examine how spatial reasoning generalizes across schema contexts. Further, the capacity limit of spatial working memory yields systematic declines in spatial reasoning as the number of objects increases [9, 52], and the distance effect [16,33] improves reasoning accuracy with larger spatial separations. Hence, our construction of VQA samples enforces fine-grained object–viewpoint distance control, and provides explicit labeling about the number of objects in each view.

Cognitive studies also revealed a hierarchical structure of spatial querying. Humans distinguish *view–object*, *view–view*, and *object–object* relations [40, 49], progressing from perceptual encoding to relational abstraction through recursive composition [13, 14]. In addition, perspective taking introduces additional cognitive complexity, engaging mechanisms of mental imagery [24, 42]. Accordingly, we differentiate queries between *self-perspective* and *perspective-change* with various types of spatial relations, to assess VLM's ability to simulate alternative views.

Analysis of spatial reasoning. Cognitive science suggests that human spatial reasoning is not a monolithic pass, but a multi-stage progression as shown in Figure 2, including the initial encoding of perceptual input, the construction of coherent mental models that integrate information across multiple views, the deductive and inferential operations to address queries, and the final decision making [18, 19]. This formulation is consistent with evidences from mental imagery and simulation [23, 24, 42]and spatial problem solving [14]. Our analysis aligns with these cognitive stages, and hence offers cognitively interpretable insights into where and how VLM reasoning diverges from human spatial cognition.

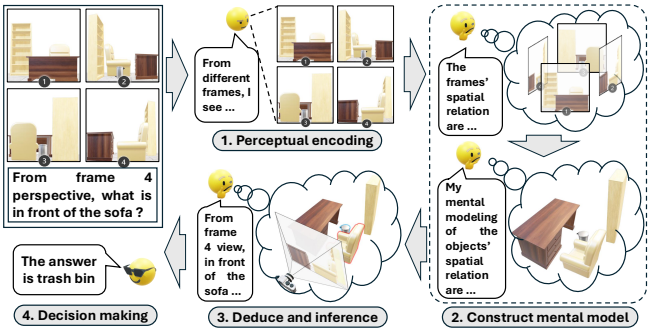

Fig. 2: Stages of humans' spatial mental modeling.

2.3 Analysis Methods

We will leverage both explicit and implicit analysis methods. Explicit analysis evaluates reasoning through generated textual rationales. We use pre-trained LLMs as graph parsers to convert VLM's reasoning paths into structured graph triplets and compare them with the ground-truth scene graphs, hence enabling structural assessment of the reasoning process beyond answer accuracy. Further, we use self-consistency prompting [50, 51] to improves robustness and faithfulness [29, 38, 51], by aggregating multiple reasoning trajectories to assess whether rationales reflect internal decision making. Yet for spatial reasoning, textual explanations often under-specify geometry or overlook occlusion and perspective constraints, potentially overstating reasoning quality.

Implicit analysis examines latent representations to uncover internal reasoning dynamics. Probing studies [32, 44, 63] test whether task-relevant structure is linearly recoverable from activations, and entropy-based analysis [22, 55, 58] tracks propagation of uncertainty to reveal the decision stability. Yet for spatial reasoning, latent activations entangle geometry with visual texture, and entropy changes signal uncertainty without localizing the cause of failure.

We integrate both manners. Explicit analysis elucidates how VLMs articulate the reasoning process, and implicit analysis reveals internal representations of

these articulations. Combining them offers fine-grained understandings of how VLMs construct and maintain spatial representations over reasoning phases.

3 Construction of VQA Samples

To facilitate our proposed analysis of VLM spatial reasoning, we construct a new spatial reasoning benchmark, namely ReMindView-Bench, which consists of >50,000 multi-view VQA samples, through a fully automated and physically grounded pipeline. This benchmark incorporates fully controlled viewpoint variations in multiple dimensions, including different spatial patterns and object relationship being queried.

3.1 Variability of Scenes and Queries

We aim to enforce two fundamental primitives of human spatial cognition, namely *relative direction* and *relative distance* [26, 49]. To systematically examine how VLMs handle these primitives in multi-view settings, we build on the cognitive factors affecting humans' spatial reasoning as described in Section 2.2 to construct carefully controlled 3D scenes and the corresponding VQA samples. The variability of constructed 3D scenes spans the following 4 spatial factors:

1. *Room type*, dining room, kitchen, bathroom, living room, and bedroom.
2. *Viewpoint spatial pattern*, view-centric and object-centric configurations.
3. *Level of distance*, fine-grained levels of distances between the camera and indoor objects, which control the scene's spatial proximity.
4. *Number of visible objects* across all rendered views.

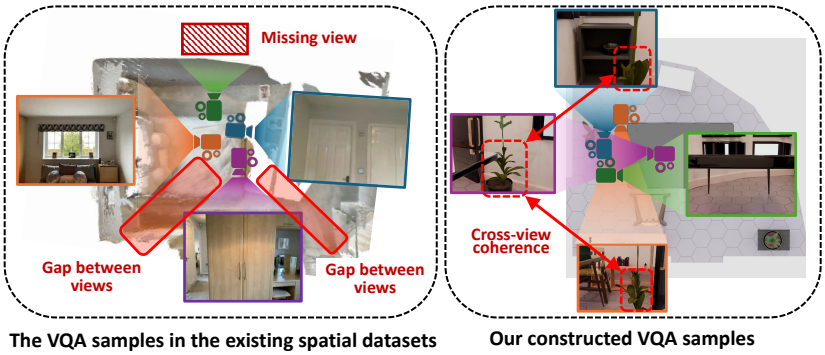

Fig. 3: Comparison of VQA samples provided by the existing spatial datasets [15,60,64] and constructed by our approach

As shown in Figure 3, by incorporating such scene variability, we can ensure cross-view coherence in VQA samples, by explicitly parameterizing complementary camera viewpoints that mathematically guarantee continuous geometric references of objects across frames. Furthermore, we ensure complete views by employing a volumetric occlusion-based visibility test and a dynamic viewpoint-searching strategy, which actively bypass occlusions to capture comprehensive spatial coverage. In contrast, existing multi-view datasets, such as

MindCube [60], SPAR-Bench [64] and 3D-CLR [15], cannot ensure such coherence because they rely on arbitrary trajectory sampling from real-world captures, which frequently results in disjointed frames containing gaps between views. They may also miss important views because of uncontrolled camera movements and natural occlusions that leave blind spots in the scene geometry.

Furthermore, variability of queries in VQA samples spans another 4 factors:

1. *Query type*, as the query concerns relative distance or relative direction.
2. *Relation type*, as the relational structure among viewpoints and objects being queried, including view–object, object–object, and view–view relations.
3. *Perspective taking*, as whether spatial reasoning is grounded in the current camera viewpoint or shifted to another object-centered viewpoint.
4. *Cross-frame reasoning*, as whether all entities involved in the query are co-observed within a single frame, or instead require multi-view integration.

We provide more details about VQA classification scheme in Appendix A.1. Some VQA samples are shown in Figure 1, and we provided more VQA examples in Appendix A.2. Statistics of the VQA samples are in Appendix A.3.

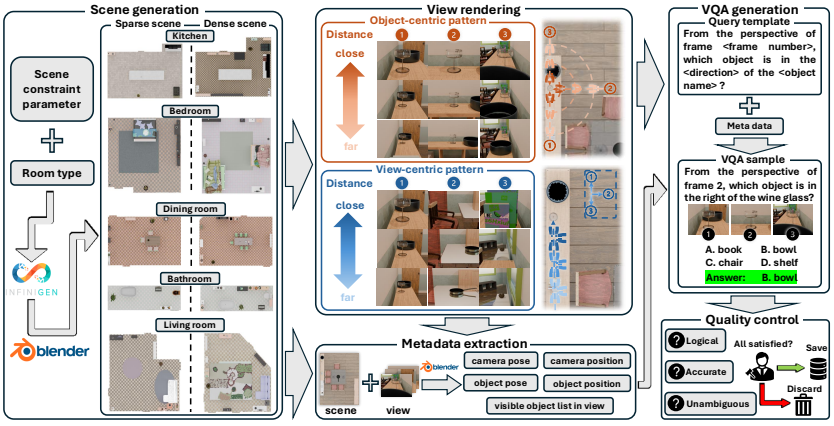

Fig. 4: Pipeline of VQA sample construction. It first generates diverse indoor 3D scenes with various room types and object densities by adjusting scene constraint parameters in Infinigen [39]. Next, multiple views are rendered in Blender with controlled camera–object distances and spatial patterns. VQA data is produced using predefined query templates, combined with metadata extracted from the scene and views.

3.2 Construction Pipeline

As shown in Figure 4, the 3D scenes and VQA samples used for our analysis are generated through a fully automated pipeline, which integrates procedural 3D scene synthesis, viewpoint-controlled rendering, structured metadata extraction, and template-based question generation.

Scene generation. We procedurally generate indoor scenes using Infinigen [39], a procedural 3D scene generator, with controllable parameters for room type, object density, and spatial layout. Detailed settings are in Appendix B.1. This

controlled generation allows systematic variation in clutter and spatial arrangement, challenging VLMs across perceptual and geometric conditions. Each 3D scene is imported into Blender [4] for rendering and metadata computation.

View rendering. For each scene, we render 4 orthogonal views for each object, following both object-centric and view-centric spatial patterns. Each pattern is captured at 10 object–viewpoint distances, regulating camera proximity and providing multi-scale spatial cues. Details of distance levels and viewpoint selection are in Appendix B.2. These settings yield structured yet diverse visual observations with core elements of spatial cognition.

Metadata extraction. Each scene–view pair is annotated with geometric metadata exported from Blender, including camera pose, object pose, and list of visible objects. It supports accurate construction of relational queries and provides geometric grounding for VQA generation.

Query generation. Using the extracted metadata, we instantiate 22 predefined query templates covering relative direction\distance, relation type and perspective-taking as shown in Figure 5. For each query, candidate answers are drawn from the visible objects in the scene, and the correct answer is computed from the metadata. An example is shown in Figure 4, with full query instantiation and answer derivation detailed in Appendix B.3.

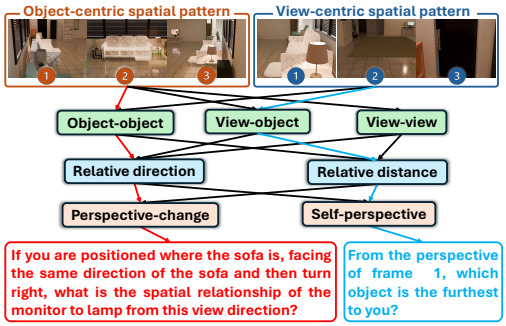

Fig. 5: Example of different query-related label combinations to generate VQ pairs.

4 VLM's Performance of Multi-View Spatial Reasoning

Based on the approach described in Section 3, we constructed >50,000 multiple-choice VQA samples that are derived from >100 procedurally generated indoor scenes. Using these VQA samples, we first evaluate the current VLMs' performances in multi-view spatial reasoning.

4.1 Evaluation Setup

VLMs being evaluated. We evaluate 15 VLMs that take multi-image input. Proprietary models include *Gemini-2.5-Pro* [10], *GPT-4o* [36] and *Claude-4-Sonnet* [1]. Open-source models include *InternVL3.5* [6], *Qwen2.5-VL* [2], *LLaVA-Video* [30] and *LLaVA-OneVision* [27] families with different parameter sizes. We also included models explicitly designed for spatial reasoning, namely *SpatialVLM* [5], *SpatialMLLM* [53] and *SpaceQwen2.5-VL* [17]. Evaluations use zero-shot settings with a unified prompting strategy (Appendix C).

Metric Design and Baselines. We consider spatial reasoning tasks as multi-choice questions (MCQs), and use the task accuracy based on exact matching as the primary evaluation metric. We provide a random-guess baseline with random selection accuracy. We further sample 600 VQA samples with 30 VQAs per query template, for human evaluators to evaluate human-level performance.

4.2 Main Results

Our evaluations reveal a pronounced gap between human spatial reasoning and current VLMs' capabilities.

Object-centric reasoning remains easier than view-centric reasoning. Figure 6 shows that nearly all VLMs performed better in object-centric spatial patterns, indicating that VLMs rely on within-view visual cues but fail to maintain relational consistency across view. Particularly, tasks involving perspective-changing or relative direction queries reveal large performance degradation, underscoring the difficulty of aligning spatial geometry and insufficient robustness to viewpoint transformations in spatial reasoning.

Viewpoint transformations amplify the reasoning instability. Figure 6 shows that VLMs have a performance drop >4% in perspective-changing settings, compared to self-perspective settings. Models that perform well under fixed viewpoints fail to preserve spatial consistency when reasoning involves perspective changing, indicating weak cross-view constancy and reference-frame alignment. From a cognitive perspective, this deficiency mirrors early-stage perceptual reasoning without mature allocentric transformation [33, 49], and shows that current attention mechanisms fail to encode cross-view correspondences, leading to disrupted spatial continuity and limited generalizability.

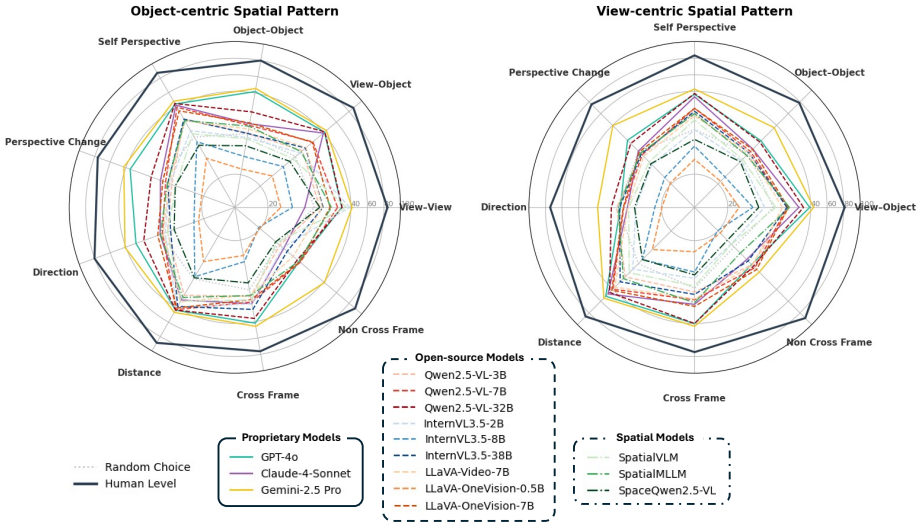

Fig. 6: Evaluation results with the 20%–40% interval expanded. More detailed results of numerical evaluations can be found in Appendix D.

Cross-frame reasoning is more challenging than within-frame reasoning. As shown in Figure 6, all VLMs exhibit a clear performance drop when reasoning requires integrating information across multiple frames. This pattern holds consistently under both object-centric and view-centric spatial patterns, highlighting the difficulty of maintaining coherent spatial representations over disjoint visual inputs. The large gap between cross-frame and non-cross-frame cases indicates that current VLMs fail to geometrically align scene elements

across complementary views, instead relying on static perceptual features in single frames. Cross-frame reasoning, which requires integrating spatial correspondences and relations, is the key to improve multi-view spatial cognition.

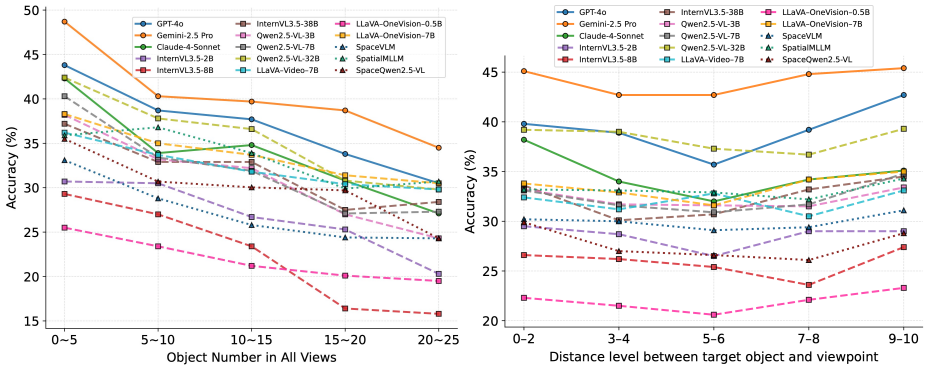

Fig. 7: Task accuracy with different numbers of objects (left) and object-viewpoint distances (right).

Number of objects and object-viewpoint distance jointly constrain spatial reasoning. As shown in Figure 7 - left, increasing the number of objects in the scene leads to accuracy degradation, which suggests that cluttered scenes amplify relational ambiguity and visual occlusion, hindering VLMs from maintaining coherent object–relation mappings. Figure 7 -right further shows that the accuracy across different object-viewpoint distances is comparatively stable. A shallow U-shaped trend is observed, where close and far distances yield higher accuracy and mid-range distances have degraded performance. This pattern suggests that extreme distances provide clearer geometric cues, either through strong visual overlap or pronounced parallax. These results demonstrate that current VLMs lack scalable mechanisms for spatial compositionality, struggling both with object-level clutter and geometric variance across views.

Real-world Validation. Furthermore, to validate that these synthetic-based findings can effectively address the sim-to-real gap and be robustly reproduced in real-world scenarios, we also conducted real-world validation over the MindCube benchmark [45, 60]. Results reported in Appendix I validated the robustness of our findings in real-world scenarios.

5 Explicit VLM Spatial Reasoning Analysis

We analyze the VLMs' spatial reasoning process with both explicit and implicit methods. First, explicit analysis examines the correctness of a VLM's observable reasoning phases and the consistency between reasoning and decision, allowing us to pinpoint where spatial reasoning succeeds or breaks across reasoning phases. The related case studies of VLMs' successes and failures are in Appendix E.1.

5.1 Analysis using LLMs as Graph Parsers

As shown in Figure 8, we design a LLM-as-graph-parser workflow to evaluate the correctness of VLM's spatial reasoning path, through a lens grounded in human

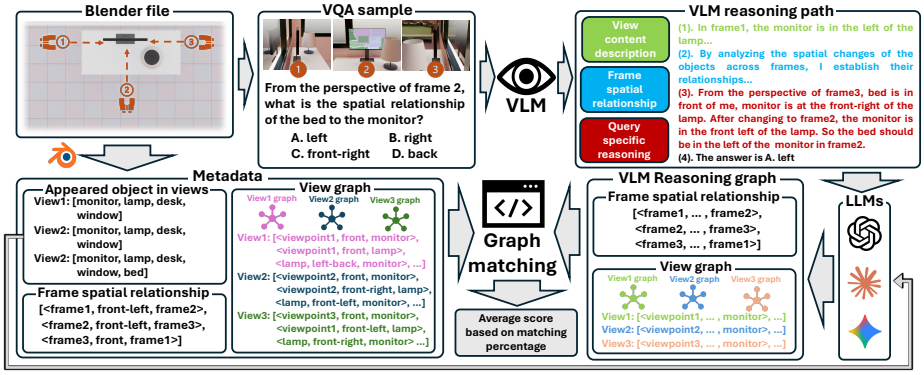

Fig. 8: LLM-as-graph-parser workflow for evaluating the VLM's spatial reasoning path

spatial cognition. The analysis extracts rich metadata from the rendered scenes in Blender, including (i) list of objects in each view, (ii) spatial relationships among viewpoints and (iii) spatial relationships between objects and viewpoints. Such metadata is then organized as separate view graphs. In each graph, nodes represent objects or viewpoints and edges encode spatial relations, so that all visible entities across views are represented within a unified relational structure.

We follow the cognitive science framework in Section 2.2 to instruct VLM's reasoning with the 4-phase template: (1) description of view content spatial relationship, (2) alignment of frame-level spatial relationship, (3) query-specific spatial reasoning, and (4) final decision. LLMs are employed to process the reasoning traces generated by VLM, with instructions to parse traces into structured graph triplets. The resulting reasoning graph is compared against the ground-truth scene graph via graph matching, to compute the proportion of correctly matched triplets. The matching score is obtained by averaging the triplet-level matching percentages across reasoning graphs. Details are in Appendix E.2.

Phase-wise reasoning reveals strong in-frame performance but degraded cross-frame inference. As shown in Table 1, our analysis exposes a distinct phase-dependent degradation pattern. To further substantiate the robustness and reliability of the LLM-based parsing procedure, we also provide results from a human evaluation of VLM reasoning paths in Appendix G.

Model	Correct (%)			Incorrect (%)		
	1	2	3	1	2	3
Qwen2.5-VL-3B	73.2	43.8	32.6	68.7	31.4	18.5
Qwen2.5-VL-7B	76.1	58.2	36.3	72.5	43.6	21.9
Qwen2.5-VL-32B	**81.9**	**63.5**	**38.8**	**76.8**	**44.2**	**24.6**

Table 1: Phase-wise reasoning correctness of Qwen2.5-VL.

The first reasoning phase achieves the highest percentage of correctness, indicating that VLMs possess strong perceptual grounding and spatial reasoning ability with *in-frame* visual contents. However, such correctness drops in later phases, reflecting a growing difficulty in maintaining geometric coherence and executing spatial inference *across views*. This trend persists for samples with incorrect final answers, where de-

scriptions remain relatively robust but later reasoning phases collapse. Larger VLMs demonstrate slower performance decay and higher phase stability, implying enhanced integration of cross-view representations. Results on other VLM families, including InternVL [6] and SpaceQwen [53] are in Appendix H.

5.2 Self-consistency Prompting

We perform self-consistency prompting by feeding multi-view inputs, reasoning path, and final answer into the same VLM to verify whether the prediction aligns with its own reasoning. This process assesses the model's self coherence and stability on spatial reasoning, revealing contradictions or drift within its reasoning trajectory. More details about prompt design are in Appendix E.3.

Self-consistency reflects the alignment between reasoning coherence and answer correctness. Results in Table 2 reveal a strong correspondence between reasoning coherence and final-answer correctness. Across all models, correct answers exhibit substantially higher self-consistency scores, indicating that successful reasoning is typically supported by internally coherent inference chains. In contrast, incorrect answers display much lower self-consistency, suggesting unstable or contradictory reasoning. This discrepancy is pronounced in small models, which may generate inconsistent rationales even when final answer is correct, reflecting shallow alignment between reasoning and decision-making. Hence, larger VLMs not only improve task accuracy but also develop more consistent reasoning-to-decision mappings, underscoring self-consistency as a robust indicator of reasoning integrity and cognitive alignment in multi-view spatial reasoning. Results on other VLM families, including InternVL [6] and SpaceQwen [53] are in Appendix H.

Model	Correct (%)	Incorrect (%)	Overall (%)
Qwen2.5-VL-3B	64.8	32.5	48.6
Qwen2.5-VL-7B	74.1	41.7	59.3
Qwen2.5-VL-32B	**83.6**	**54.2**	**68.4**

Table 2: Self-consistency prompting of Qwen2.5-VL family.

6 Implicit VLM Spatial Reasoning Analysis

Implicit analysis complements the explicit analysis, by probing latent representations of VLMs such as entropy distribution and representation pattern. While explicit analysis measures what a VLM claims to reason, implicit analysis reveals how its internal representations evolve and deteriorate over reasoning phases.

6.1 Linear Probing Analysis

We conduct linear probing analysis by extracting token logits from the end of each reasoning phase, along with the final answer logits from all QA samples. A 4-layer MLP probe is then trained using cross-entropy loss, and analysis is performed by computing the loss between probe's predicted logits and VLM's answer logits on the test set. This analysis provides a quantitative measure of how VLM's latent representations degrade across successive spatial reasoning phases.

VLM representations exhibit progressive degradation of task-relevant information over reasoning phases and model sizes. As shown in Figure 9, linear probing results reveal a consistent increase in cross-entropy loss when the reasoning progresses or the model enlarges. Early-phase representations exhibit the lowest loss, indicating that perceptual encoding maintains stronger alignment with the final decision signal. As reasoning progresses, the loss rises steadily, i.e., spatial relations are transformed into higher-level and distributed representations that are less aligned with output space. Although larger models achieve higher accuracy, their latent features become increasingly abstract, resulting in higher probe loss. The divergence between correct and incorrect samples also grows in later phases, suggesting that successful reasoning depends on preserving stable information flow from perceptual encoding to decision making, and failed reasoning reflects cumulative degradation of task-relevant representations.

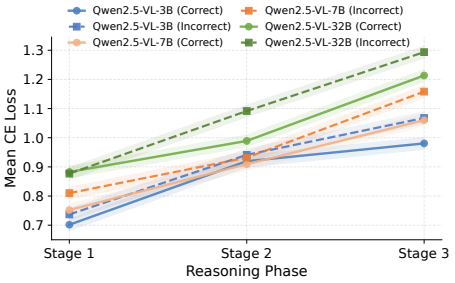

Fig. 9: Linear probing analysis on the Qwen2.5-VL family.

The type of query relationship reveals distinct degradation patterns in latent representations. To examine how relational structure affects representation stability, we stratify probing results by query type, including *view-view*, *view-object*, and *object-object*. These categories impose different spatial reasoning requirements and exhibit distinct probing dynamics.

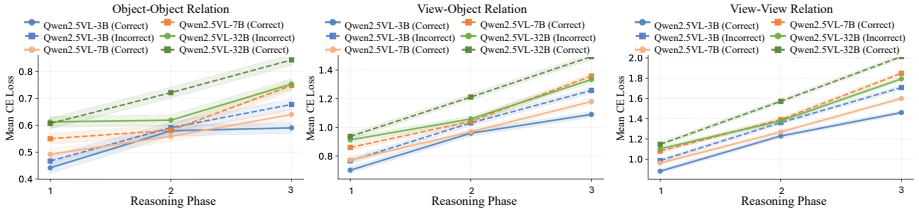

Fig. 10: Linear probing analysis over different types of query relationship

As shown in Figure 10, object-object queries show the lowest cross-entropy loss and the smallest separation between correct and incorrect samples across reasoning phases, suggesting that object-centric spatial relations remain relatively stable and linearly accessible in the latent representation. In contrast, view-view queries exhibit the highest loss and the steepest phase-wise increase, with the largest divergence between correct and incorrect cases, indicating that reasoning across viewpoints requires more complex spatial transformations that progressively degrade representational alignment with the final decision space. View-object queries display intermediate behavior, with moderate loss growth and separation between correct and incorrect samples.

Overall, degradation of task-relevant information is influenced by relation types in the query. Object-object reasoning preserves stable representations, and

viewpoint-related relations introduce more abstraction and representational instability across reasoning phases, indicating that multi-view spatial transformations progressively weaken the linear accessibility of task-relevant information.

6.2 Entropy Dynamics Analysis

We analyze the evolution of entropy by computing the mean and variance of token-level logit entropy at each reasoning phase. Details are in Appendix F.

Model confidence distinctly separates final answer correctness in reasoning phases. As shown in Figure 11, correct and incorrect reasoning cases exhibit a consistent and well-defined separation in entropy. For all models, correct reasoning maintains substantially lower entropy, indicating more confident reasoning trajectories. In contrast, incorrect reasoning display persistently higher entropy, reflecting uncertain or conflicting internal representations. This separation is most pronounced in larger models, where correct reasoning paths are both low-entropy and tightly distributed, indicating calibrated confidence, while incorrect paths remain diffuse. This finding indicates that entropy patterns strongly modulate spatial reasoning correctness, and entropy is an effective proxy for diagnosing reliability and correctness in spatial reasoning.

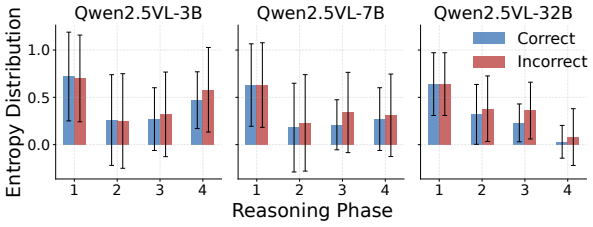

Fig. 11: Entropy distribution of Qwen2.5-VL familiy.

Different spatial reasoning factors produce distinct entropy dynamics. We further analyze entropy trajectories under different spatial VQA factors, including cross-frame effects, spatial organization, and perspective change.

As shown in Figure 12, for cross-frame reasoning, entropy increases more noticeably in later reasoning phases compared to non-cross-frame queries. This suggests that uncertainty mainly arises during the integration of spatial information across view-

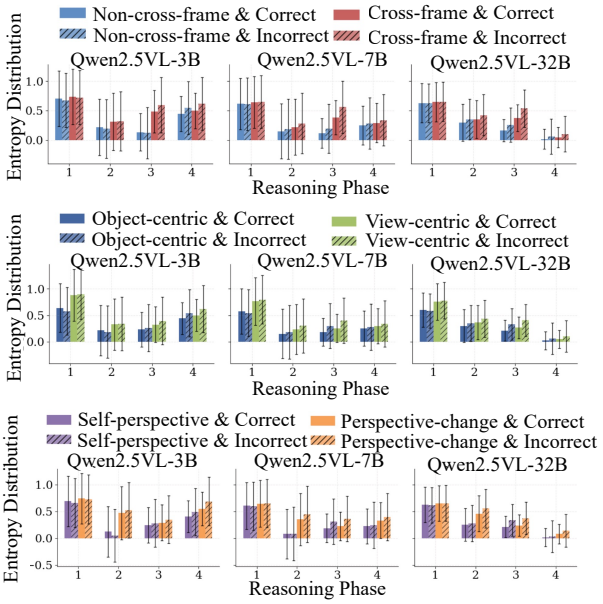

Fig. 12: Entropy distribution over different VQA factors.

points, where maintaining cross-view geometric consistency becomes challenging. In contrast, view-centric queries exhibit higher entropy in early reasoning phases, while the difference gradually diminishes later. This indicates that the main difficulty lies in establishing a stable viewpoint-dependent reference frame during the initial stages of reasoning. Perspective-change queries display a different pattern, with a sharp entropy increase in the intermediate reasoning phase. This spike suggests that reference-frame transformation required for perspective taking introduces large uncertainty, which propagates into subsequent steps.

Overall, these entropy dynamics show that spatial factors stress different reasoning stages: cross-frame queries affect late-stage integration, view-centric queries raise early viewpoint uncertainty, and perspective changes introduce mid-stage instability.

7 Conclusion of Findings

We present a cognitively grounded framework for analyzing VLM's spatial reasoning in multi-view settings, instantiated through carefully controlled 3D scenes and VQA samples. By combining explicit reasoning path evaluation and implicit latent representation probing, we showed fine-grained evidences about where and how current VLMs diverge from human-like spatial cognition. Their limitations arise mainly from insufficient cross-view alignment, unstable inference progression and poor confidence calibration across reasoning phases.

Our analyses provide complementary insights from both explicit and implicit perspectives. In *explicit analysis*, evaluation of VLM's reasoning paths reveals a gap between perceptual grounding and geometric inference. Models remain accurate in early perceptual encoding phase but degrade in relational alignment and cross-view reasoning, and self-consistency prompting confirms that stable reasoning chains correlate with correct outcomes. In *implicit analysis*, linear probing reveals a degradation of task-relevant spatial information across reasoning phases, i.e., intermediate representations are more abstract and less aligned with the final decision signal. This degradation varies across query types, as object–object relations have stable representations, and viewpoint-related relations exhibit higher instability. Entropy dynamics further show a separation between correct and incorrect reasoning trajectories: correct cases have lower entropy and incorrect cases display higher uncertainty.

Together, the primary limitation of current VLM spatial reasoning is not perception but maintaining coherent spatial representations across reasoning phases and viewpoints. Improvements hence requires preserving cross-view geometric structure and maintaining stable information flow in reasoning.

Acknowledgments

We thank the reviewers and the area chair for their insightful comments and feedback. This work was supported in part by National Science Foundation (NSF) under grant number IIS-2205360 and CCF-2215042, and National Institutes of Health (NIH) under grant number R01HL170368 and R01HL184168.

References

1. Anthropic: Claude 4 sonnet: Advanced language and vision model. https://www.anthropic.com/news/claude-3-5-family (2024), accessed: 2025-11-09
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987)
4. Blender Foundation: Blender (2025), https://www.blender.org, open-source 3D creation suite
5. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
8. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025)
9. Clevenger, P.E., Hummel, J.E.: Working memory for relations among objects. Attention, Perception, & Psychophysics **76**(7), 1933–1953 (2014)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2018)
12. Freksa, C.: Spatial and temporal structures in cognitive processes. In: Foundations of computer science: potential—theory—cognition, pp. 379–387. Springer (2006)
13. Gentner, D.: Structure-mapping: A theoretical framework for analogy. In: Cognitive Science, vol. 7, pp. 155–170. Wiley Online Library (1983)
14. Hegarty, M.: Mechanical reasoning by mental simulation. Trends in Cognitive Sciences **8**(6), 280–285 (2004)
15. Hong, Y., Lin, C., Du, Y., Chen, Z., Tenenbaum, J.B., Gan, C.: 3d concept learning and reasoning from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9202–9212 (2023)
16. Huttenlocher, J., Hedges, L.V., Duncan, S.: Categories and particulars: prototype effects in estimating spatial location. Psychological review **98**(3), 352 (1991)
17. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135 (2025)
18. Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press (1983)
19. Johnson-Laird, P.N., Byrne, R.M.J.: Deduction. Lawrence Erlbaum Associates (1991)

20. Kahneman, D., Treisman, A., Gibbs, B.J.: The reviewing of object files: Object-specific integration of information. Cognitive psychology **24**(2), 175–219 (1992)
21. Kamath, A., Hessel, J., Chang, K.W.: What's" up" with vision-language models? investigating their struggle with spatial reasoning. arXiv preprint arXiv:2310.19785 (2023)
22. Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., Gal, Y.: Semantic entropy probes: Robust and cheap hallucination detection in llms. arXiv preprint arXiv:2406.15927 (2024)
23. Kosslyn, S.M.: Image and Mind. Harvard University Press (1980)
24. Kosslyn, S.M.: Image and Brain: The Resolution of the Imagery Debate. MIT Press (1994)
25. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954 (2020)
26. Levinson, S.C.: Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press (2003)
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
28. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. arXiv preprint arXiv:2505.21500 (2025)
29. Li, J., Yan, H., He, Y.: Drift: Enhancing llm faithfulness in rationale generation via dual-reward probabilistic inference. In: The 63rd Annual Meeting of the Association for Computational Linguistics: ACL 2025 (2025)
30. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
31. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
32. Liu, Z., Zhu, T., Tan, C., Lu, H., Liu, B., Chen, W.: Probing language models for pre-training data detection. arXiv preprint arXiv:2406.01333 (2024)
33. Loomis, J.M., Da Silva, J.A., Fujita, N., Fukusima, S.S.: Visual space perception and visually directed action. Journal of experimental psychology: Human Perception and Performance **18**(4), 906 (1992)
34. Ma, C., Lu, K., Cheng, T.Y., Trigoni, N., Markham, A.: Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. Advances in neural information processing systems **37**, 68803–68832 (2024)
35. Ogezi, M., Shi, F.: Spare: Enhancing spatial reasoning in vision-language models with synthetic data. arXiv preprint arXiv:2504.20648 (2025)
36. OpenAI: Gpt-4o: Multimodal large language model. `https://openai.com/research/gpt-4o` (2024), accessed: 2025-11-09
37. Ouyang, K.: Spatial-r1: Enhancing mllms in video spatial reasoning. arXiv e-prints pp. arXiv–2504 (2025)
38. Parcalabescu, L., Frank, A.: On measuring faithfulness or self-consistency of natural language explanations. arXiv preprint arXiv:2311.07466 (2023)
39. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., et al.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21783–21794 (2024)

40. Rieser, J.J.: Access to knowledge of spatial structure at novel points of observation. Journal of Experimental Psychology: Learning, Memory, and Cognition **15**(6), 1157 (1989)
41. Scholl, B.J.: Objects and attention: The state of the art. Cognition **80**(1-2), 1–46 (2001)
42. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science **171**(3972), 701–703 (1971)
43. Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., Vosoughi, S.: Judging the judges: A systematic study of position bias in llm-as-a-judge. arXiv preprint arXiv:2406.07791 (2024)
44. Stańczak, K., Hennigen, L.T., Williams, A., Cotterell, R., Augenstein, I.: A latent-variable model for intrinsic probing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13591–13599 (2023)
45. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
46. Tarr, M.J., Pinker, S.: Mental rotation and orientation-dependence in shape recognition. Cognitive psychology **21**(2), 233–282 (1989)
47. Tenderra, R.M., Theves, S.: Human intelligence relates to neural measures of cognitive map formation. Cell Reports **44**(8) (2025)
48. Trivedi, P., Gulati, A., Molenschot, O., Rajeev, M.A., Ramamurthy, R., Stevens, K., Chaudhery, T.S., Jambholkar, J., Zou, J., Rajani, N.: Self-rationalization improves llm as a fine-grained judge. arXiv preprint arXiv:2410.05495 (2024)
49. Tversky, B.: Functional, cognitive, and embodied aspects of spatial thinking. Proceedings of the International Conference on Spatial Cognition **1404**, 1–10 (2005)
50. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
51. Wang, Y., Atanasova, P.: Self-critique and refinement for faithful natural language explanations. arXiv preprint arXiv:2505.22823 (2025)
52. Wong, J.H., Peterson, M.S., Thompson, J.C.: Visual working memory capacity for objects from different categories: A face-specific maintenance effect. Cognition **108**(3), 719–731 (2008)
53. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
54. Xiang, S., Yang, A., Xue, Y., Yang, Y., Feng, C.: Self-supervised spatial reasoning on multi-view line drawings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12745–12754 (2022)
55. Xiong, M., Santilli, A., Kirchhof, M., Golinski, A., Williamson, S.: Efficient and effective uncertainty quantification for llms. In: Neurips Safe Generative AI Workshop 2024 (2024)
56. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
57. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
58. Ye, Z., Melo, L.C., Kaddar, Y., Blunsom, P., Staton, S., Gal, Y.: Uncertainty-aware step-wise verification with generative reward models. arXiv preprint arXiv:2502.11250 (2025)

59. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
60. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
61. Yu, L., Chen, X., Gkioxari, G., Bansal, M., Berg, T.L., Batra, D.: Multi-target embodied question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6309–6318 (2019)
62. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905 (2025)
63. Zhang, A., Chen, Y., Pan, J., Zhao, C., Panda, A., Li, J., He, H.: Reasoning models know when they're right: Probing hidden states for self-verification. arXiv preprint arXiv:2504.05419 (2025)
64. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
65. Zhang, W., Zhou, Z., Zheng, Z., Gao, C., Cui, J., Li, Y., Chen, X., Zhang, X.P.: Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. arXiv preprint arXiv:2503.11094 (2025)
66. Zhang, Z., Wu, Y., Jia, L., Wang, Y., Zhang, Z., Li, Y., Ran, B., Zhang, F., Sun, Z., Yin, Z., et al.: Think3d: Thinking with space for spatial reasoning. arXiv preprint arXiv:2601.13029 (2026)
67. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625 (2025)