Diagnosing Spatial Consistency across Perspectives and Viewpoints in Large Vision-Language Models

Yoonji Kim[1], Jieun Kim[2], Yujin Jeong[1], Sung-Bae Cho[1]
[1]Dept. of Computer Science, Yonsei University, Seoul, South Korea
[2]Dept. of Artificial Intelligence, Yonsei University, Seoul, South Korea
{yoonjikim,lilly9928,yujinj00,sbcho}@yonsei.ac.kr

Abstract

Consistent reasoning about 3D spatial relations across changing viewpoints is fundamental for Embodied AI agents operating in dynamic environments. While Large Vision-Language Models (LVLMs) have advanced multimodal perception, their ability to maintain spatial consistency across diverse perspectives remains underexplored. Existing benchmarks primarily assess spatial capabilities from a static, single-view, and egocentric perspective, failing to capture the dynamic nature of real-world spatial cognition. To address this gap, we introduce **SCOPE** (**S**patial **CO**nsistency across **PE**rspectives and Viewpoints), a comprehensive benchmark designed to rigorously diagnose spatial reasoning capabilities. Grounded in human cognitive theories of dual spatial representations, SCOPE discretizes the 360° field into multiview scenarios to systematically evaluate both allocentric and egocentric reasoning capabilities. Our dataset comprises **20.1K** spatial VQA pairs derived from high-quality 3D environments. Through an extensive evaluation of 26 state-of-the-art LVLMs, we identify two fundamental limitations that prevent consistent spatial understanding across viewpoints. We hope **SCOPE** facilitates the diagnosis of spatial reasoning, serving as a stepping stone toward reliable embodied action.

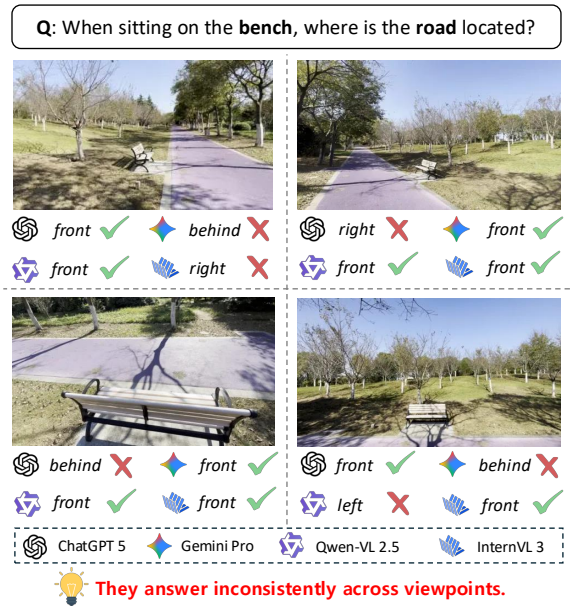


Figure 1: **Spatial Inconsistency across Viewpoints.** Although the spatial relationship remains invariant in the physical world, LVLMs produce contradictory answers across different viewpoints. This suggests that current models struggle to achieve a consistent understanding of 3D space.

1 Introduction

The ability to derive coherent 3D spatial understanding from viewpoint-dependent observations is an essential prerequisite for Embodied AI operating in dynamic and multi-view environments (Wang et al., 2024; Zhu et al., 2025a). Currently, Large Vision-Language Models (LVLMs) enable robust multimodal perception and reasoning (Zhu et al., 2025b; Chen et al., 2025; Bai et al., 2025; Grattafiori et al., 2024; Reid et al., 2024), and serve as the basis for emerging Vision-Language-Action (VLA) models (Kim et al., 2025; Li et al., 2024b; Physical Intelligence et al., 2025).

However, as illustrated in Figure 1, LVLMs often produce inconsistent spatial judgments across viewpoints, even under identical geometry. Such viewpoint-induced inconsistency limits their ability to reliably comprehend and interact with dynamic real-world environments.

Despite this critical requirement, existing research primarily evaluates the spatial reasoning capabilities of LVLMs from a single-view and egocentric (camera-centered) perspective (Cheng et al., 2024; Kamath et al., 2023; Li et al., 2024a). Consequently, the extent to which LVLMs integrate allocentric (object-centered) perspective and multi-view information to consistently infer 3D spatial

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32803–32827
July 2-7, 2026

relations remains largely underexplored.

To bridge this gap, we draw on human spatial cognition, which maintains a robust understanding of the dynamic environment. Mirroring classic cognitive experiments designed to assess this ability, we present diagnosing **S**patial **CO**nsistency across **PE**rspectives and Viewpoints benchmark (**SCOPE**). **SCOPE** comprises three tasks that target different aspects of allocentric and egocentric reasoning under both single-view and multi-view settings. For each scene, we discretize the 360° field around the camera at 45° intervals, capturing eight viewpoints and constructing spatial queries.

Our dataset comprises **20,144** spatial VQA pairs across **7,512** views from **939** real-world and synthetic scenes, designed to evaluate whether models can integrate multi-view observations into a unified and consistent spatial understanding. Furthermore, we propose diagnostic metrics to assess viewpoint-dependent spatial reasoning. Together with our tasks, these metrics allow us to characterize model behavior under viewpoint changes in terms of consistency, robustness, and common failure patterns.

Interestingly, in our spatial integration task, we use simple few-shot prompting that mimics human spatial cognition: first grounding relations in an egocentric frame and then refining them in an allocentric frame. This improves the model's spatial judgments, yielding an average gain of **6.4%**.

Our key contributions are as follows:

- **Cognitively Grounded Benchmark.** We introduce **SCOPE**, a 360° benchmark that systematically evaluates LVLMs' spatial consistency through three tasks derived from human spatial cognition experiments.
- **Extensive Experiments.** We demonstrate that 1) scaling improves egocentric but not allocentric spatial reasoning, 2) LVLMs significantly lag behind humans in spatial updating and cross-view integration, and 3) spatial fine-tuning yields task-specific gains without broad generalization.
- **Cognitively Inspired Improvement.** We show that guiding LVLMs to follow human-like spatial strategies via simple few-shot prompting leads to improved spatial understanding.

2 Related Work

Spatial Reasoning in Human Cognition. Cognitive science has long shown that robust spatial reasoning relies on multiple complementary abilities. Developmental research highlights Level-2 perspective-taking across viewpoints, as measured by Piaget's "three mountains" task (Piaget et al., 1957), supporting *relational*, *depth*, and *orientation* reasoning. Classic spatial-updating experiments by Simons & Wang (Simons and Wang, 1998) show that self-motion supports maintaining a coherent representation of space, capturing spatial updating under observer motion and grounding *relational*, *depth*, *orientation*, and *occlusion* reasoning. Finally, studies of object continuity, such as object files (Kahneman et al., 1992) and amodal completion (Kellman and Spelke, 1983), explain how humans preserve stable object representations despite occlusion, incomplete observations, or ambiguous depth cues, reflecting *relational*, *depth*, and *occlusion* reasoning. However, these classical paradigms typically assume a shared and stable environment. In contrast, LVLMs may receive only partial observations, motivating benchmarks that explicitly test scene coherence: whether observations are integrated into a consistent spatial representation.

Benchmarks for Spatial Reasoning in LVLMs. Prior studies generally utilized mixed perspectives without a clear distinction. For instance, VSR (Liu et al., 2023) employs mixed perspective but lacks an explicit distinction between egocentric and allocentric relations, leading to potential ambiguity. Recent works attempt to explicitly separate these perspectives but are largely limited to single, fixed perspectives. Specifically, What's Up (Kamath et al., 2023) and SpatialRGPT (Cheng et al., 2024) focus primarily on egocentric (camera-based) view, making it difficult to assess allocentric reasoning, while TopViewrs (Li et al., 2024a) restricts evaluation to a bird's-eye View. Although subsequent studies such as 3DSRBench (Ma et al., 2025a) and Spatial-MM (Shiri et al., 2024) explicitly distinguish both perspectives and incorporate allocentric relations using a depicted agent's facing direction or 3D pose (object-intrinsic axes), they remain largely confined to single-view images. As a result, correct answers may reflect spurious correlations or shortcut reasoning rather than grounded 3D understanding. We therefore evaluate models based on the consistency of their predictions across all viewpoints of the same scene.

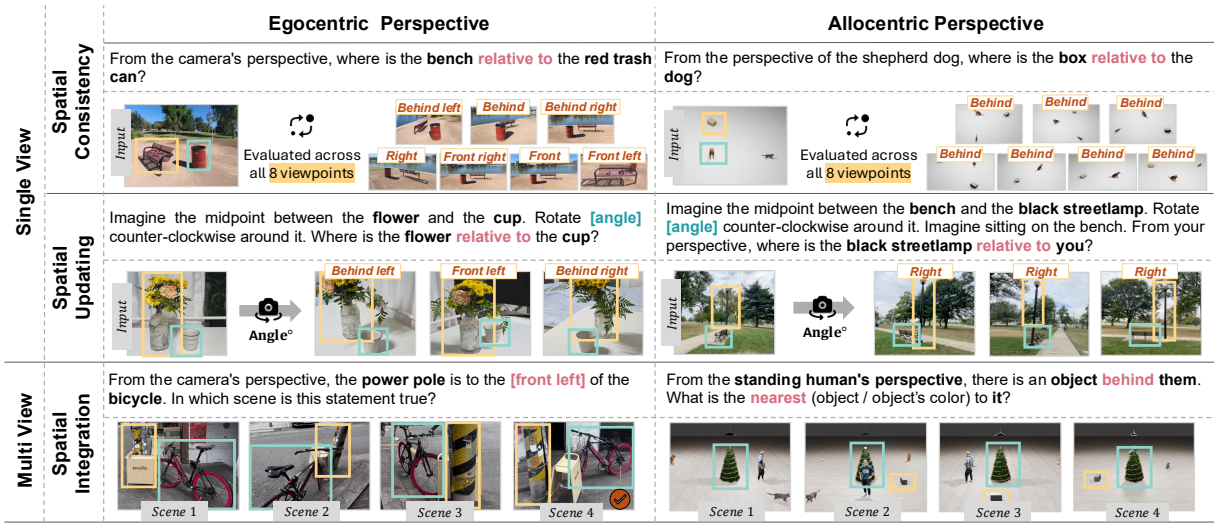


Figure 2: **Overview of SCOPE.** The figure illustrates the tasks in SCOPE. All tasks query spatial relations between a target object and a reference object.

3 SCOPE

We propose **SCOPE**, a benchmark that evaluates spatial reasoning in dynamic 3D environments, with an emphasis on maintaining consistent spatial understanding across viewpoint changes. As illustrated in Figure 2, SCOPE comprises three tasks. Each task consists of 8 viewpoints spanning a full 360° rotation at 45° increments.

3.1 Task Taxonomy

As shown in Table 1, SCOPE decomposes spatial reasoning into cognitively grounded factors. We identify four core factors that characterize the complexity of spatial reasoning: **Relational (Rel.)** captures spatial relations between objects and serves as the primary signal of 3D understanding: whether a model can determine where an object is in the scene. **Depth (Dep.)** requires reasoning about relative distances between objects in 3D space, which is essential for disambiguating relations that appear similar in 2D projections. **Orientation (Ori.)** requires identifying a reference object within the scene and inferring its facing direction to anchor spatial judgments. **Occlusion (Occ.)** tests whether models can infer visibility constraints: determining what is hidden or revealed from a given viewpoint. In our setting, Relational and Depth form the base spatial queries, while Orientation and Occlusion introduce additional 3D challenges.

Building on these query types, we design tasks organized into three groups to test whether models can consistently infer spatial relations within the same space. **Spatial Consistency** tests whether models maintain stable spatial representations across viewpoint changes. **Spatial Integration** requires aggregating partial observations from multiple views to infer global scene structure. **Spatial Updating** evaluates the ability to perform mental rotation under imagined viewpoint shifts.

Group	Perspective	Factors			
		Rel.	Dep.	Ori.	Occ.
Spatial Consistency	Egocentric	✓	✓		
	Allocentric	✓	✓	✓	
Spatial Integration	Egocentric	✓	✓		
	Allocentric	✓	✓	✓	✓
Spatial Updating	Egocentric	✓	✓		✓
	Allocentric	✓	✓	✓	✓

Table 1: **Factor coverage of task groups.** Rel.: Relational, Dep.: Depth, Ori.: Orientation, Occ.: Occlusion.

3.2 Benchmark Construction

We construct SCOPE through a two-stage pipeline (Figure 3). Each item is represented as $(I_{n,\theta}, Q_{n,\theta}, O_{n,\theta}, y_{n,\theta})$, where I, Q, O, and y denote the image, question, options, and ground-truth label, respectively. Here, n indexes the scene and $\theta \in \{0°, 45°, \ldots, 315°\}$ denotes the viewpoint angle. Starting from a base item at $\theta = 0°$, we generate variations by rotating the observer counterclockwise around the scene center. See Appendix A for additional details on data construction.

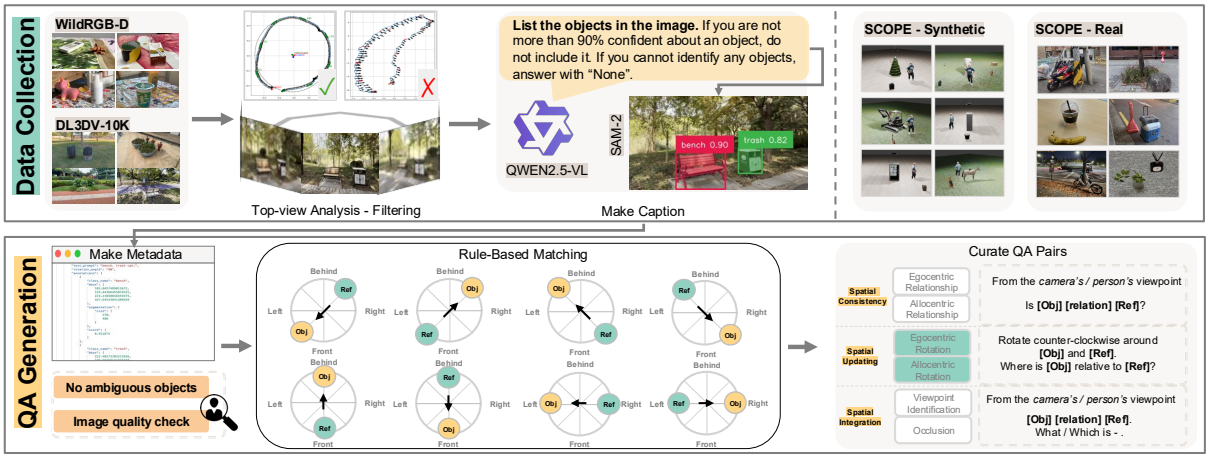


Figure 3: **Data Generation Pipeline for SCOPE.** Top: image filtering and top-view analysis. Bottom: rule-based QA pair generation tests spatial consistency across 8 directions.

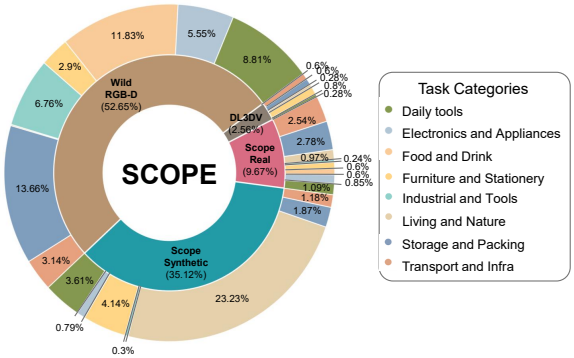


Figure 4: **Composition of the SCOPE benchmark.** Inner ring: distribution of data sources. Outer ring: task categories by object types.

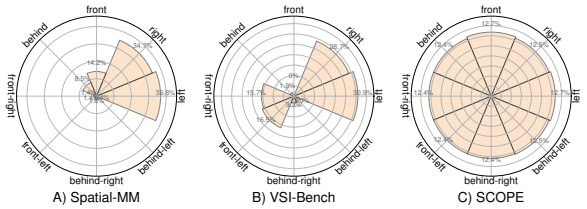


Figure 5: **Directional distribution of questions across benchmarks.** Each plot shows the distribution of questions across eight directions.

As shown in Figure 4, SCOPE is carefully constructed to capture diverse spatial relationships in varying contexts of the scene. To mitigate directional biases, we balance the directional distribution of QA pairs (Figure 5).

Data Collection. We collect data from three sources: existing 360° datasets, synthetic rendering, and expert-curated photography. For real-world data, we leverage DL3DV-10K (Ling et al., 2024) and WildRGB-D (Xia et al., 2024). We filter images based on camera pose accuracy using top-view visualization, remove small or heavily occluded objects to avoid part-whole ambiguity (Varzi, 2007), and exclude scenes with significant vertical displacement. For synthetic data, we render indoor and outdoor scenes in Blender (Hess, 2013) by uniformly sampling 37 object models. We also collaborate with experts to capture real-world photographs with diverse spatial configurations.

Question-Answer Generation. For question construction, we generate captions for DL3DV and WildRGB-D scenes using the Qwen2.5-72B (Bai et al., 2025). All captions are reviewed and validated by three domain experts. We then apply Grounded-SAM2 (Kirillov et al., 2023) to obtain instance masks and compute object centroids, which are also reviewed by the same experts. For synthetic rendering and expert-curated scenes, we directly use object-level spatial metadata. Finally, spatial questions are generated using task-specific rule-based templates.

4 Experiments

4.1 Experimental Setup

Evaluated Models. We evaluate a comprehensive set of models on SCOPE, grouped into five categories: (1) **Baselines**: Random (25% for 4-way multiple choice), Frequency (selecting the most frequent answer), and a text-only Blind GPT-5.2 baseline; (2) **Open-source models**: Qwen2.5-VL (Bai et al., 2025), Llama-3.2-Vision (Li et al., 2025), InternVL2.5 (Zhu et al., 2025b), LLaVA-OneVision (Grattafiori et al., 2024), and

Method	Rank	Avg	Spatial-Consistency		Spatial-Updating		Spatial-Integration	
			Ego	Allo	Ego	Allo	Ego	Allo
Human	-	91.24	97.66	83.12	88.67	90.00	94.67	93.33
Baseline								
Random	-	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Frequency	-	25.56	25.00	25.38	26.16	25.00	25.38	26.43
GPT5.2-Blind	-	22.28	23.84	27.19	25.26	22.94	22.48	11.96
Open-Source Models								
Qwen2.5-VL-3B	9	33.71	46.37	25.79	24.25	24.49	25.14	56.20
Qwen2.5-VL-7B	8	34.52	58.51	24.90	20.68	25.06	26.08	51.88
Qwen2.5-VL-32B	5	36.23	64.36	25.38	24.93	16.28	25.73	60.71
Qwen2.5-VL-72B	3	37.78	68.23	25.48	26.75	24.30	33.99	47.95
Llama-3.2-11B	13	29.82	35.87	26.24	21.49	25.03	25.57	44.69
InternVL2.5-2B	12	30.26	44.32	25.00	20.99	24.65	25.00	41.61
InternVL2.5-8B	7	35.22	67.03	25.06	25.17	25.41	23.44	45.18
InternVL2.5-14B	2	38.83	74.86	24.46	20.50	25.10	26.27	61.76
InternVL2.5-38B	1	39.98	75.52	25.82	26.49	24.78	24.50	62.77
InternVL2.5-78B	4	36.47	78.94	25.51	26.82	25.82	27.10	34.64
LLaVA-OV-4B	10	32.08	56.16	26.11	24.72	26.90	26.82	31.79
LLaVA-OV-8B	6	35.56	57.05	25.10	24.88	25.63	24.53	56.16
Gemma3-4B	11	30.36	25.00	22.59	24.88	25.38	24.50	59.82
Gemma3-12B	14	28.91	25.00	25.13	26.44	24.11	21.70	51.07
Gemma3-27B	15	28.44	25.00	13.20	25.83	26.40	27.00	53.21
Proprietary Models								
GPT-5.2	5	40.68	66.11	26.71	28.57	25.00	50.75	46.96
GPT-5-nano	4	42.67	75.45	30.01	29.65	24.45	46.25	50.18
GPT-5-mini	1	58.55	89.55	34.90	48.45	26.24	83.56	68.57
Gemini-2.5-Pro	2	51.26	88.89	22.40	47.37	26.19	25.00	97.68
Gemini-2.5-Flash	3	50.05	85.40	25.92	46.78	25.26	22.29	94.64
Gemini-2.5-Flash-Lite	6	32.26	32.43	25.49	24.88	24.08	25.40	61.25
Spatial-Specialized Models								
SpatialRGPT	4	34.07	58.14	25.22	22.84	23.10	25.00	50.09
SpatialReasoner	5	31.51	47.74	26.43	22.55	25.82	24.59	41.94
RoboBrain-3B	3	35.49	71.06	26.33	16.45	24.62	26.18	48.30
RoboBrain-7B	2	40.92	81.49	25.76	17.39	26.68	25.28	68.93
RoboBrain-32B	1	45.37	87.43	25.51	23.73	24.75	28.63	82.14

Table 2: **Evaluation on SCOPE.** Zero-shot setting with sampling temperature set to 0. Orange and yellow indicate the best and second-best results, respectively, within each section and column.

Gemma3 (Gemma Team et al., 2025); (3) **Proprietary models**: GPT-5 and Gemini-2.5 (Comanici et al., 2025); (4) **Spatial-specialized models**: SpatialRGPT (Cheng et al., 2024), SpatialReasoner (Ma et al., 2025b), and RoboBrain (Team et al., 2025); and (5) **Human Evaluation**: a human performance upper bound used as a reference (details in Appendix B.1).

4.2 Main Results

Table 2 presents results on **SCOPE**. To diagnose the gap between humans and LVLMs, our analysis reveals four consistent patterns.

Scaling does not resolve allocentric consistency. Across model families, scaling up improves egocentric spatial consistency, but it does not reliably fix allocentric spatial consistency. For example, Qwen2.5-VL scales from 3B to 72B while allocentric consistency remains at chance (3B: 25.79 → 72B: 25.48). A similar trend appears in InternVL2.5 (2B: 25.00 → 78B: 25.51). Even proprietary models show inconsistent gains: GPT-5-mini reaches 34.90, whereas Gemini-2.5-Pro drops to 22.40, both far below human performance (83.12). Notably, the blind baseline achieves 27.19 on allocentric consistency, indicating that larger vision-language capacity alone does not guarantee robust allocentric viewpoint generalization.

Spatial updating is brittle even within the same 3D space. Our results show that LVLMs struggle not only with allocentric perspective-taking but also with egocentric viewpoint changes, indicating limited spatial updating. In egocentric-updating, most open-source models remain near chance (e.g., Qwen2.5-VL-7B: 20.68; InternVL2.5-14B: 20.50), with the best reaching only 26.82 (InternVL2.5-

78B). Even top proprietary models are still far below human performance (88.67). The gap widens further in allocentric-updating, where all models remain essentially at chance (e.g., GPT-5-mini: 26.24; Gemini-2.5-Pro: 26.19; best overall: 26.90 with LLaVA-OV-4B) compared to humans (90.00). Together, these results suggest that models fail to maintain a coherent spatial representation and consequently struggle to reason from perspectives beyond the current view.

Multi-view integration breaks at relational composition. Our results suggest that the primary difficulty in multi-view spatial reasoning lies not in aggregating views, but in composing relations across them. Many LVLMs achieve near-ceiling performance on allocentric integration while remaining close to chance on egocentric integration (except GPT models): Gemini-2.5-Pro scores 97.68 on Spatial-Integration (Allo) but only 25.00 on Spatial-Integration (Ego). Even under weak allocentric consistency, this pattern suggests that in multi-view settings, models can answer allocentric queries by matching entities across views while relations remain fixed, without constructing a coherent allocentric map, whereas egocentric integration exposes the failure to compose relations that vary across viewpoints.

Fine-tuning yields limited generalization. Spatial-specialized (fine-tuned) models improve specific metrics but do not generalize uniformly across SCOPE. For example, RoboBrain-32B attains high egocentric consistency (87.43) and strong integration (Allocentric: 82.14), yet remains near chance level on allocentric consistency (25.51) and shows weak spatial updating (Egocentric: 23.73; Allocentric: 24.75). Similarly, RoboBrain-7B reaches 81.49 on egocentric consistency but only 17.39 on egocentric-updating and 26.68 on allocentric-updating. This indicates that boosting performance on isolated spatial subtasks is insufficient; robust spatial intelligence likely requires jointly improving data diversity, viewpoint supervision, and integration-centric objectives.

4.3 Further Analysis

Consistency Analysis. We quantify multi-view agreement using the following consistency score:

$$C(\mathcal{S}) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left[\bigwedge_{v\in\mathcal{S}} T_{v\to v_0}\big(f(I_{n,v})\big) = f(I_{n,v_0})\right] \quad (1)$$

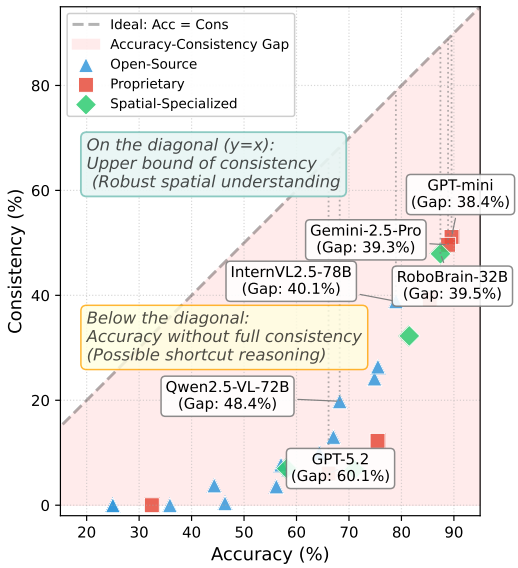


Figure 6: **Accuracy vs. Consistency.** We plot accuracy against the consistency score (Eq. 1). The shaded region above the diagonal indicates accuracy inflation attributed to shortcut reasoning.

Here, $f(I_{n,v})$ is the prediction for scene n at viewpoint v, and $T_{v\to v_0}$ maps predictions from v to the reference frame v_0.

Figure 6 reveals a substantial gap between accuracy and consistency across all model categories. Even top-performing models like GPT-mini (89.6% accuracy) and Gemini-2.5-Pro (88.9% accuracy) achieve only around 50% consistency, indicating that roughly half of their correct answers fail to generalize across viewpoint variations. Notably, spatial-specialized models do not escape this pattern—RoboBrain-32B achieves 87.4% accuracy but only 47.9% consistency, suggesting that even models explicitly trained for spatial reasoning rely heavily on viewpoint-specific cues. In contrast, human participants achieve 87.5% consistency, far exceeding all models (Appendix C.4). Overall, the accuracy–consistency gap suggests reliance on shortcuts rather than robust spatial reasoning, motivating evaluation of consistency alongside accuracy.

Relational Bias Analysis. Figure 7 exposes two consistent tendencies across model families and scales. First, many models exhibit a clear diagonal weakness: errors concentrate more heavily on diagonal confusions than on orthogonal confusions, suggesting that composing heading and depth cues jointly is harder than resolving a single axis at a time. Second, when restricting attention to orthog-

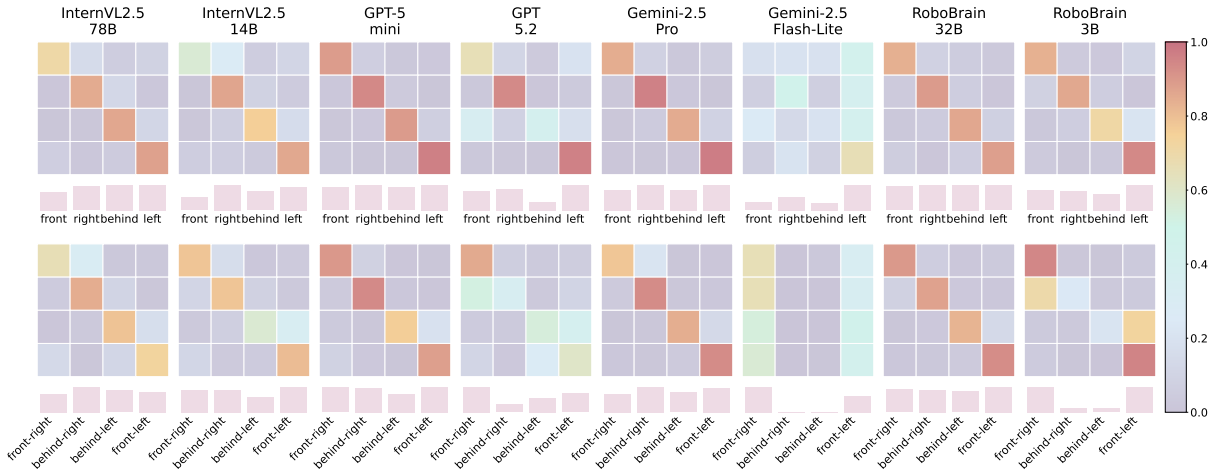


Figure 7: **Relational confusion matrices for egocentric spatial reasoning.** Each column pairs a high-performing model (left) with its weak-performing model (right) within the same model family. The top row shows orthogonal (front, right, behind, left), while the bottom row shows diagonal (front-left, front-right, behind-right, behind-left).

onal pairs, models generally discriminate left/right more accurately than front/behind, indicating that these models successfully acquire the horizontal coordinate system but struggle with depth perception and heading estimation. Beyond accuracy, Figure 7 also reveals a qualitative failure mode in weaker models: despite a balanced query distribution over relations, their predictions collapse toward a limited subset of relations (i.e., they almost never output certain labels), evidencing a strong relational bias. Together, these findings suggest that current models fail to capture the full geometric structure of spatial relations and instead rely on incomplete and biased relational cues.

Finding 1: *LVLMs spatial reasoning is non-uniform, with predictions often collapsing toward biased relations.*

Error Analysis. Most models struggle with spatial updating. To better understand these failures, we conduct a detailed error analysis. Figure 8(a) reveals a highly consistent relation-wise profile: performance peaks at right, whereas front-left, front, and front-right are uniformly the most challenging. The fact that this trend holds across diverse architectures and scales suggests that the dominant difficulty is spurious correlations that systematically favor certain relations and make others harder to infer under spatial updating. Figure 8(b) further breaks down the error types by condition (diagonal vs. orthogonal) and rotation angle, categorizing failures into No Rotation, Wrong Direction (clockwise rotation with the correct angle), and Wrong Angle (counterclockwise rotation with an incorrect angle). Across both orthogonal and diagonal settings, models struggle more under 90° and 270° rotations than under 180°, indicating that quarter-turn updates pose a greater challenge than half-turn updates. Notably, the dominant failure mode at 90° and 270° is Wrong Direction, implying that models often apply rotation in the incorrect direction.

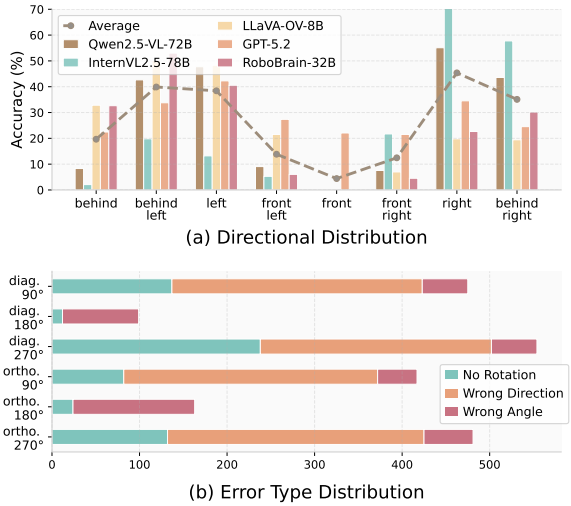


Figure 8: **Error analysis on Spatial Updating.** (a) Predicted relation distribution after spatial updating across models. (b) Error type distribution for the best-performing model (GPT-5-mini).

Finding 2: *LVLMs fail at spatial updating, often confusing front-oriented relations and wrong rotation directions.*

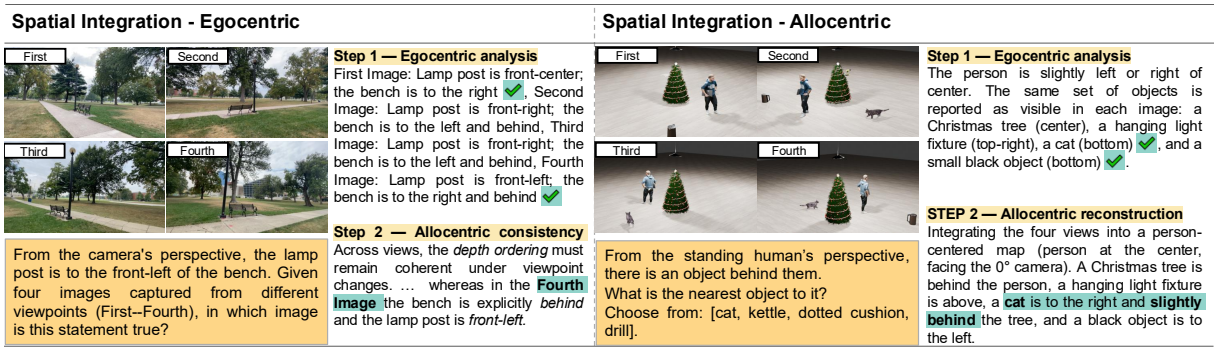


Figure 9: **Case studies of Egocentric → Allocentric prompting.** Allocentric reconstruction resolves ambiguities that remain under per-view egocentric cues and supports consistent multi-view spatial inference under occlusion.

Model	Egocentric			Allocentric		
	Orig.	Mimics	Gap	Orig.	Mimics	Gap
LLaVA-OV-8B	24.53	28.50	+3.97	56.16	62.70	+6.54
InternVL2.5-8B	23.44	30.10	+6.66	45.18	47.41	+2.23
Gemma-3-4B	24.50	30.60	+6.10	59.82	**65.90**	+6.08
Llama-3.2-11B	25.57	**34.24**	+8.67	44.69	55.70	+11.01
Average	24.51	30.86	+6.35	51.46	57.93	+6.47

Table 3: **Spatial integration performance** (accuracy, %). Yellow highlights the best result per task.

Cognitively Inspired Prompting. Human spatial cognition operates through a two-stage process: initially grounding spatial relations egocentrically (from one's own viewpoint), then integrating these into an allocentric frame (object-to-object relations) (Hao et al., 2017; Levinson, 2003). Drawing inspiration from this cognitive strategy, we redesign our prompting approach to mirror this natural progression. Full prompts are provided in Appendix B.2. Table 3 shows that mimicking an Egocentric → Allocentric reasoning procedure consistently improves spatial integration across all four models. On Egocentric, accuracy increases from 24.51% to 30.86% on average (+6.35), with the largest gain from Llama-3.2-11B (+8.67). On Allocentric, the average improvement is similarly strong (+6.47; 51.46% → 57.93%), driven primarily by a large jump for Llama-3.2-11B (+11.01; 44.69% → 55.70%). Overall, these results suggest that guiding models to follow human-like Ego-to-Allo integration yields more reliable multi-view spatial reasoning.

Case Studies. Figure 9 illustrates why an explicit Egocentric → Allocentric procedure improves spatial reasoning. In the egocentric spatial integration task, a per-view inspection based on local cues can be underconstrained: two views appear plausible because they share similar lateral configurations, making it difficult to choose using a single snapshot. The allocentric step resolves this by integrating observations across views into a unified scene layout and enforcing cross-view consistency, which eliminates the spurious candidate and selects the correct view. In the allocentric spatial integration task under occlusion, the egocentric step first enumerates objects visible around the standing person from each viewpoint, while the allocentric reconstruction consolidates these partial observations into a single map centered on the person. This global map makes it straightforward to infer what lies behind the person and to identify the nearest object among the provided choices, demonstrating how allocentric reconstruction provides a robust constraint beyond view-specific cues.

5 Concluding Remarks

In this paper, we introduce SCOPE, a benchmark of 20.1K QA pairs designed to evaluate LVLMs' 3D spatial cognition through tasks grounded in human spatial cognition research. SCOPE evaluates three progressive tasks: Spatial Consistency, Spatial Updating, and Spatial Integration. Our evaluation across 26 state-of-the-art models shows that (i) LVLMs' spatial reasoning is non-uniform, with predictions collapsing toward biased relations, (ii) spatial updating is brittle even within the same 3D space, and (iii) spatial fine-tuning yields task-specific gains without broad generalization. We further find that a simple human-inspired prompting strategy improves consistency, yielding an average gain of 6.4%. We hope SCOPE serves as a practical diagnostic tool and a stronger target for future LVLM research on building world models consistent across perspectives and viewpoints.

6 Limitations

Our benchmark provides a controlled and consistent evaluation of model performance under a unified experimental setting; however, several limitations remain. All models were tested using a single fixed prompting configuration with deterministic decoding (sampling temperature = 0). While this design improves comparability across models, performance may vary under alternative prompting strategies or decoding settings. In addition, because all evaluation questions were written in English, the extent to which our findings generalize to other languages remains to be validated.

7 Ethical Considerations

Our benchmark comprises data from publicly available 360° datasets as well as synthetic samples generated through direct rendering. The publicly sourced datasets contain photorealistic scenes, and the synthetic data are rendered in a manner that does not introduce copyright issues. Moreover, our data sources include no personal data, no uniquely identifiable individuals, and no offensive or harmful content.

Acknowledgments

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv:2502.13923*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv:2501.17811*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 135062–135093.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv:2503.19786*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv:2407.21783*.

Xin Hao, Yi Huang, Yiying Song, Xiangzhen Kong, and Jia Liu. 2017. Experience with the cardinal coordinate system contributes to the precision of cognitive maps. *Frontiers in Psychology*, 8.

Roland Hess. 2013. *Blender Foundations: The Essential Guide to Learning Blender 2.5*. Routledge, London, U.K.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. 1992. The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2):175–219.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9161–9175.

Philip J. Kellman and Elizabeth S. Spelke. 1983. Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15(4):483–524.

Jieun Kim and Sung-Bae Cho. 2026. Neuro-symbolic reasoning with multiple large language models combined by first-order logic. In *Hybrid Artificial Intelligent Systems*, pages 227–238. Springer Nature Switzerland.

Jieun Kim, Yujin Jeong, and Sung-Bae Cho. 2026. Visual–linguistic abductive reasoning with LLMs for knowledge-based visual question answering. In *Findings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 6529–6544.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, and 1 others. 2025. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 2679–2713.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213.

Stephen C. Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge Univ. Press, Cambridge, U.K.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research (TMLR)*.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024a. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1786–1807.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, and 1 others. 2024b. CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv:2411.19650*.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, and 1 others. 2024. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics (TACL)*, 11:635–651.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025a. 3DSRBench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6924–6934.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. 2025b. SpatialReasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, and 1 others. 2025. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of the Conference on Robot Learning (CoRL)*.

Jean Piaget, Bärbel Inhelder, F. J. Langdon, and J. L. Lunzer. 1957. The child's conception of space. *British Journal of Educational Studies*, 5(2):187–189.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21440–21455.

Daniel J. Simons and Ranxiao Frances Wang. 1998. Perceiving real-world viewpoint changes. *Psychological Science*, 9(4):315–320.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, and 1 others. 2025. RoboBrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Achille C. Varzi. 2007. Spatial reasoning and ontology: Parts, wholes, and locations. In *Handbook of Spatial Logics*, pages 945–1038. Springer.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, and 1 others. 2024. EmbodiedScan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19757–19767.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. 2024. RGBD objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22378–22389.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643.

Ruihan Yang, Haoyu Chen, Jiayuan Zhang, Mingyu Zhao, Cheng Qian, Kangrui Wang, Qiang Wang, Tejaswini Koripella, Mohammad Movahedi, Mingyu Li, Heng Ji, Hanwang Zhang, and Tianmin Zhang.

2025b. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Eric Zhao, Vikash Raval, He Zhang, Jiayuan Mao, Shaohan Shangguan, Stefanos Nikolaidis, Yuke Wang, and Daniel Seita. 2025. ManipBench: Benchmarking vision-language models for low-level robot manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*.

Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. 2025a. SPA: 3d spatial-awareness enables effective embodied representation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025b. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2508.18265*.

Table of Contents

A Construction and Annotation Pipeline

We utilize the validation splits of DL3DV-10K (Ling et al., 2024) and WildRGB-D (Xia et al., 2024), supplemented with professionally captured real scenes and Blender-generated synthetic scenes. After filtering, we retain 10 scenes from DL3DV-10K, 223 scenes from WildRGB-D, 40 curated real scenes, and 664 synthetic scenes, for a total of 939 unique scenes. For each scene, we render 8 viewpoints, resulting in 7,512 views in total.

A.1 Data Sources and Privacy

DL3DV-10K. A large-scale dataset featuring 51.2M frames from 10,510 videos across diverse point-of-interest locations. All videos include COLMAP-calculated camera poses in precise 4×4 transformation matrices, enabling accurate trajectory extraction and 360° coverage verification.

WildRGB-D. Comprises 8,500 objects across 46 categories with nearly 20,000 RGB-D videos captured in real-world cluttered environments. Direct depth acquisition enables accurate 3D annotations, and videos include camera pose annotations and reconstructed point clouds for precise object centroid computation.

Blender Synthetic Scenes. For synthetic scenes, we use only 3D assets with licenses permitting commercial use (CC0, CC-BY). The dataset covers 37 distinct object categories in total.

Privacy and Content. All scenes contain only natural objects and everyday items without personally identifiable information or human subjects.

A.2 Camera Pose and Top-View

To construct viewpoint variations $(I_{n,\theta}, Q_{n,\theta}, O_{n,\theta}, y_{n,\theta})$ for $\theta \in \{0, 45, \ldots, 315\}$, we require scenes with complete 360° circular camera trajectories at precise 45° intervals. We develop a rigorous filtering pipeline based on geometric analysis of camera pose metadata.

Camera Pose Metadata Structure. Each image is associated with a 4×4 transformation matrix encoding camera pose in world coordinates:

$$T = \begin{bmatrix} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where $R \in \mathbb{R}^{3\times3}$ encodes camera orientation and $\mathbf{t} = (t_x, t_y, t_z)^\top$ encodes camera position.

Trajectory Extraction and Scene Center Determination. We extract camera positions $(x, y, z) = (t_x, t_y, t_z)$ and project them onto the horizontal plane for trajectory analysis. To determine the scene center and verify circular motion, we fit a circle to the projected positions $\{(x_i, y_i)\}_{i=1}^N$ using least-squares optimization. We retain only scenes where the mean fitting residual is below 0.1 meters, ensuring geometric consistency. The fitted center (c_x, c_y) serves as the pivot point for computing angular positions.

Camera orientation vectors (u_x, u_y) are extracted from the rotation matrix by projecting the third column of R (the camera's forward direction) onto the horizontal plane, yielding $(u_x, u_y) = (R_{13}, R_{23})$ after normalization. This enables verification that cameras face toward the scene center.

Angular Discretization and Coverage Verification. For each camera position, we compute its angular position relative to the scene center: $\phi_i = \text{atan2}(y_i - c_y, x_i - c_x)$. We partition the 360° viewing space into eight angular bins centered at target orientations $\{0, 45, 90, 135, 180, 225, 270, 315\}$ corresponding to cardinal and intercardinal directions. Scenes are retained only if they provide camera views within ± 11.25 of all eight target angles, ensuring complete directional coverage consistent with our automatic verification criterion.

Vertical Displacement Filtering. To ensure unambiguous horizontal spatial relations, we keep only frames whose camera height stays close to that of the first image. Let $z_i = t_z$ be the camera height of image i, and compute the standard deviation across the scene as

$$\sigma_z = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2}.$$

We then retain only images satisfying

$$|z_i - z_1| \leq \max(0.5\,\text{m},\, 1.5\sigma_z),$$

and discard the rest.

Top-View Visualization. We generate top-view visualizations plotting camera positions (x, y) with orientation arrows derived from (u_x, u_y). These visualizations enable manual verification that cameras form consistent circular trajectories with uniform angular spacing and inward-facing orientations toward scene centers, confirming geometric

Caption-Generation Prompts

DL3DV-10K

```
"Name two main objects in this image shortly."
```

WILD-RGBD

```
"Tell me the names of the objects on the tablecloth or blanket. There are usually 1--4 objects. If you are not more than 90% confident about an object, do not include it. If you cannot identify any objects, answer with 'None'. To help your judgment, I will provide a representative description: {class_}. Answer with numbers like 1., 2. and separate each item with a new line."
```

Figure 10: **Caption-generation prompts.** Text instructions used to obtain object names for DL3DV-10K and WILD-RGBD before segmentation.

validity for generating multi-view spatial reasoning questions.

A.3 Prompt Templates

Our benchmark construction pipeline relies on a two-stage prompting approach: first, we use a vision-language model to identify candidate objects in each scene, and then we leverage these predictions to guide automated segmentation.

Caption Generation Prompts. To obtain candidate object names for each image, we query Qwen2.5-VL-72B with dataset-specific prompts designed to elicit concise, high-confidence object labels. The prompts differ slightly between datasets to account for their distinct visual characteristics: DL3DV-10K contains diverse outdoor and indoor scenes with salient objects, while WILD-RGBD focuses on tabletop arrangements with small objects placed on textured surfaces. Figure 10 shows the exact prompts used for each dataset. For DL3DV-10K, we use a simple open-ended prompt asking the model to name two main objects. For WILD-RGBD, we provide more explicit instructions emphasizing confidence thresholds and formatting constraints to ensure reliable object identification in cluttered tabletop scenes.

SAM2 Input and Output Format. After obtaining open-vocabulary object names from Qwen2.5-VL-72B, we construct structured inputs to SAM2 to generate precise segmentation masks (Figure 11). Specifically, SAM2 takes (i) the RGB image and (ii) a text prompt listing candidate object categories,

SAM2 Input and Output Format

Input

`Image:` raw RGB image for a given scene.
`Text prompt:` list of object names predicted by Qwen2.5-VL-72B, concatenated with periods.

Output

`class_name:` predicted object label for each segmented instance.
`bbox:` 2D bounding box coordinates in image space.
`segmentation:` pixel-wise segmentation mask (stored as RLE).
`score:` confidence score for the prediction. 1

Figure 11: **SAM2 input–output format.** Structured representation of the image and text inputs and the resulting class, box, mask outputs used in our benchmark.

and returns instance masks (and corresponding bounding boxes) for the queried objects. We then use these masks to compute object centroids and spatial extents, which serve as the geometric foundation for generating and automatically verifying spatial-relation questions in SCOPE.

A.4 Data Filtering Statistics

To ensure caption quality, we filtered out unsuitable top-view scenes and images with fewer than two identifiable objects. The process consists of automated filtering followed by manual review, and Table 5 summarizes the number of scenes retained at each stage.

A.5 Human Annotation and Validation

After filtering, three trained annotators manually reviewed all remaining images and captions, correcting errors and removing scenes with unclear or ambiguous objects. All annotators were research team members receiving no additional compensation beyond their employment.

A.6 Directional Rules

We implement rules covering the diagonal and axis-aligned configurations summarized in Table 4. Crucially, the ground-truth $y_{n,\theta}$ differs by reference frame: allocentric labels remain fixed across θ (object relations are view-invariant), while egocentric labels vary with θ (relations depend on camera perspective).

Template Generation Process. Given a scene with object centroids extracted from instance masks, we: (1) determine the initial egocentric

0°	45°	90°	135°	180°	225°	270°	315°	360°
behind-right	right	front-right	front	front-left	left	behind-left	behind	behind-right
front-left	left	behind-left	behind	behind-right	right	front-right	front	front-left
front-right	front	front-left	left	behind-left	behind	behind-right	right	front-right
behind-left	behind	behind-right	right	front-right	front	front-left	left	behind-left
front	front-left	left	behind-left	behind	behind-right	right	front-right	front
behind	behind-right	right	front-right	front	front-left	left	behind-left	behind
right	front-right	front	front-left	left	behind-left	behind	behind-right	right
left	behind-left	behind	behind-right	right	front-right	front	front-left	left

Table 4: **Directional transformation rules across camera rotations.** Eight representative patterns showing directions transform as the camera rotates counter-clockwise in 45° increments from 0° to 360°.

Source	Original	Filter	Review
DL3DV-10K	10,510	147	10
WildRGB-D	∼20,000	268	223
Total	–	415	233

Table 5: **Scene filtering statistics.** Number of scenes retained after each filtering stage. Final 233 scenes yield 1,864 viewpoints.

placement $(d_{\text{Obj}}, d_{\text{Ref}})$ from the first viewpoint, (2) select the corresponding transformation rule from the eight basic patterns, (3) compute the rotation angle θ for each additional viewpoint using camera pose metadata, and (4) derive the ground-truth direction from the rule's mapping. We then instantiate question templates with these predicted spatial relationships, yielding questions that are approximately uniformly distributed over the eight horizontal reloations.

Option Construction. Each question is accompanied by four answer options. For directional tasks, distractors are drawn from the same group of orthogonal or diagonal directions, never mixing across groups (e.g., front vs. front-left). For spatial integration tasks, all options are carefully curated to be clearly distinguishable from one another. These choices ensure that model performance reflects genuine spatial reasoning.

B Experiment Details

B.1 Human Evaluation Details

We conduct a human evaluation on a 188-instance subset of SCOPE, denoted **SCOPE-188**. This subset contains 8 scenes per task and is constructed by stratified sampling over rotation angles and spatial relation types to closely match the distribution of the full benchmark.

Participants and Setup. We recruited thirteen human participants, all holding at least a bachelor's degree. Participants were compensated with $10 USD for their participation. The evaluation was administered through a web-based interface (Figure 12) that displays one or four images, the corresponding instruction, and four answer options (e.g., A/B/C/D). Each SCOPE-188 instance is *identical* to the multiple-choice question used in the LVLM evaluation, with the same task formulation, prompt, and answer options; we only use a smaller subset for human testing.

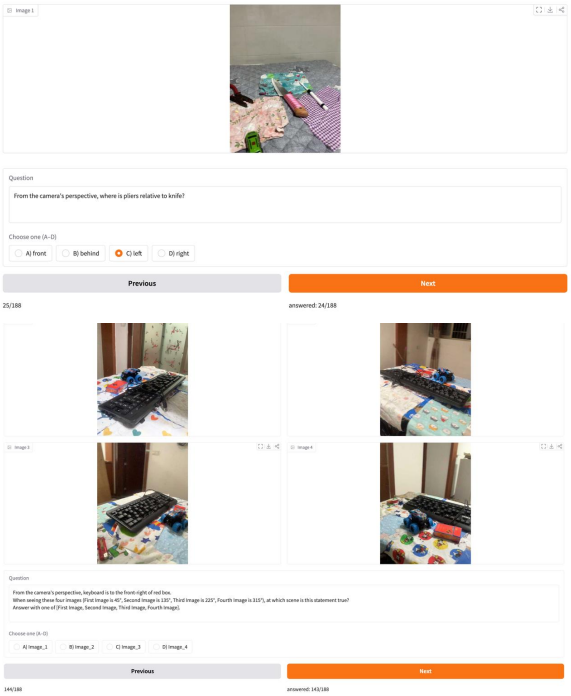


Figure 12: **Human evaluation interface.** Web-based UI used by human participants for SCOPE-188.

B.2 Task-specific LVLM Prompts

For all three tasks in SCOPE, we use a shared prompting protocol across all evaluated LVLMs. Each model receives the same visual input and question text; we only adapt a short suffix to constrain the answer space. Thus, the prompts are fully deterministic and identical across models. Fig-

ure 13 summarizes the text prompts that are appended to the task-specific questions.

Ego-to-Allocentric Reasoning. To examine whether LVLMs can mimic human-like reasoning, we use an explicit ego-to-allocentric instruction that decomposes the task into two stages: (i) **egocentric** *multi-view* integration, where the model aggregates evidence across multiple camera views in an observer-centered reference frame (prompt in Figure 14), and (ii) **allocentric** *multi-view* integration, where the model aligns evidence across views into a shared scene-centered reference frame (prompt in Figure 15). In the Mimics setting, the model decides at which viewing angle a given relation holds.

Task-specific LVLM Prompts

Spatial Consistency and Spatial Updating

Single image + question q + options o (e.g., [Front, Right, Behind, Left]).

```
"Answer with exactly one of: A, B, C, D."
```

Spatial Integration - Egocentric

Four-view images + question q + options o (e.g., [First Image (0°), ...]).

```
"Answer with exactly one of: First Image,
Second Image, Third Image, Fourth Image."
```

Spatial Integration - Allocentric

Four-view images + question q + options o (e.g., [Cat, Tree, Trumpet, Horse]).

```
"Answer with exactly one of: A, B, C, D."
```

Figure 13: **Task-specific LVLM prompts.** Text suffixes appended to questions for the three SCOPE tasks.

Spatial Integration(egocentric) Prompts

You will see four images of the same 3D scene taken from different camera viewpoints. Given a statement "{statement}" about the spatial relation between the *subject* ({subject}) and the *reference* ({reference}), your task is to select the image in which the statement is true. Answer with exactly one of: First Image, Second Image, Third Image, Fourth Image.

Step 1 – Egocentric view (per-image, camera-centered).
For each image, identify the subject and reference objects, then judge their relation directly from the camera view. Use on-screen cues to infer depth (front/behind) and lateral position (left/right), and combine them to infer diagonal relations (e.g., front-left). Record whether the statement holds for each image.

Step 2 – Allocentric view (scene-level integration), if needed.
If egocentric judgments leave multiple plausible candidates, integrate evidence from all four views into a single coherent scene layout. Use cross-view consistency checks to eliminate spurious candidates and determine which view uniquely satisfies the statement.

Step 3 – Answer.
Return exactly one option: First Image / Second Image / Third Image / Fourth Image.

Figure 14: **Ego-to-Allocentric reasoning prompt for Spatial Integration (Egocentric).**

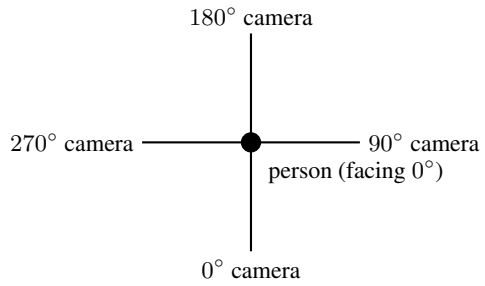
Spatial Integration(allocentric) Prompts

Ego-to-Allocentric Reasoning Prompt (Task6).
You will see four images of the same 3D scene taken from different camera viewpoints (angle_info). A person stands at the center of the scene and faces the 0° camera. Given the question "{question}" and choices choices, answer with exactly one label: A/B/C/D.

Step 1 – Egocentric observations (per-image, camera-centered).
For each image, locate the standing person and list the objects visible around them (especially those related to the choices). From the current camera view, describe what appears in front of the person (between person and camera), behind the person (far side from the camera), to the left, and to the right. Note that objects behind the person may be occluded in some views and visible in other views; record these observations per image.

Step 2 – Allocentric reconstruction (bird's-eye, person-centered).
Integrate the four views into a single person-centered map. Place the person at the center, facing the 0° direction, and define directions from the person's perspective: front $\rightarrow 0^\circ$, behind $\rightarrow 180^\circ$, left $\rightarrow 270^\circ$, right $\rightarrow 90^\circ$. Place all observed objects into this unified layout, using occlusion logic and cross-view consistency to resolve missing or conflicting evidence.

Mental Map (from above):



Step 3 – Answer.
Using the reconstructed map, identify the object behind the person and determine which choice is nearest to it. Output the final answer with exactly one label: A/B/C/D.

Figure 15: **Ego-to-Allocentric reasoning prompt for Task6 (Allocentric Integration).**

Model	Δ
GPT-4o-mini	+0.71%
RoboBrain-32B	-1.56%
Qwen-72B	+3.96%
InternVL-38B	+1.52%
Gemma-12B	-1.35%
LLaVA-8B	+1.55%
LLaMA-11B	+0.32%
Average	+1.01%

Table 6: **Performance change under natural language paraphrasing.** Δ denotes task-averaged accuracy change relative to original template questions.

System Prompt for Natural Command Generation

Natural Language Rewriting for Embodied AI

```
"You are helping create natural commands/questions for an embodied AI spatial reasoning benchmark. Your job is to rewrite a template-based question into a single natural command or question that a real person would say to a robot or embodied AI agent."
```

Figure 16: **System prompt for natural command generation.** Instruction used to convert template-based questions into natural user-like commands for embodied AI agents.

C Additional Experiments

C.1 Prompt Robustness Analysis

We conduct an additional experiment to assess whether model performance is driven by template-pattern gaming rather than genuine spatial reasoning. Specifically, we paraphrase all 20.1K template questions into natural, conversational phrases using GPT-5.2 with the following system prompt.

The task-averaged performance change relative to the original template questions (Δ) for representative models is shown in Table 6.

Since SCOPE isolates spatial reasoning with minimum visual ambiguity and linguistic confounds, natural paraphrasing does not degrade performance ($|\Delta| \leq 4\%$ across all models), indicating that results are unlikely to be driven by template-pattern gaming.

C.2 Directional Distribution Analysis

To examine directional bias in existing benchmarks, we compared the distribution of directional expressions across Spatial-MM (Shiri et al., 2024), VSI-Bench (Yang et al., 2025a), and SCOPE. All expressions were mapped to eight canonical directions (left, right, front, behind, front-right, front-left, behind-right, behind-left).

For Spatial-MM, we used the Spatial-Obj subset with 814 directional multiple-choice questions (39 excluded due to ambiguous mapping). For VSI-Bench, we used 968 multiple-choice questions specifying relative directions.

As shown in Figure 5, both existing benchmarks exhibit significant imbalance—most references concentrate on left–right relations, with other axes underrepresented. In contrast, SCOPE provides balanced coverage across all spatial directions.

C.3 Analysis Relational matrices

Figure 17, which extends the confusion matrices shown in Figure 7, presents relational confusion matrices where rows correspond to the ground-truth directions and columns correspond to the model predictions. The color intensity of each cell represents the proportion of samples with a given ground-truth direction that are classified into a particular predicted direction. Consequently, darker red diagonal cells indicate a higher proportion of correct predictions, while darker red off-diagonal cells reveal systematic confusions between specific directional pairs.

From the figure, proprietary models that are not extremely small generally achieve strong performance. Among open-source models, large InternVL variants and RoboBrain exhibit relatively low directional bias, whereas many other models show pronounced biases, such as consistently failing to predict certain directions or over-preferring specific ones.

C.4 Human Consistency Analysis

We report human performance on **SCOPE-188** to assess whether the accuracy–consistency gap reflects genuine model limitations rather than inherent task difficulty (Table 7).

Humans achieve 87.50% full-scene consistency, confirming that consistent spatial understanding across all eight views is achievable. The large human–model gap (59.23%) suggests that low model consistency reflects genuine spatial reasoning failures rather than single-view perceptual difficulty.

	Accuracy	Orth (4-view)	Diag (4-view)	All (8-view)
Human ($n = 13$)	97.72%	98.56%	96.88%	87.50%
LVLM (top 13 models)	77.69%	49.97%	40.71%	28.27%
Gap	20.03%	48.59%	56.17%	59.23%

Table 7: **Human vs. LVLM performance on SCOPE-188.**

Correlation	ManipBench	EB-N + EB-M
Pearson	0.88 ($p = 0.0007$)	0.81 ($p = 0.05$)
Spearman	0.98 ($p \approx 0$)	0.83 ($p = 0.04$)
Kendall	0.91 ($p \approx 0$)	0.73 ($p = 0.05$)

Table 8: **Correlation between SCOPE and embodied benchmarks.** ManipBench: 10 overlapping models. EmbodiedBench (EB-N + EB-M): 6 overlapping models.

Task	Pearson	Spearman
Spatial Consistency (Ego)	-0.04	0.30
Spatial Consistency (Allo)	-0.07	-0.06
Spatial Updating (Ego)	-0.06	-0.25
Spatial Updating (Allo)	0.08	0.12
Spatial Integration (Ego)	0.12	0.07
Spatial Integration (Allo)	-0.25	-0.34

Table 9: **Human–model error correlation on SCOPE-188.** Near-zero correlations indicate that model failures stem from reasoning rather than perceptual ambiguity.

C.5 Embodied AI Correlation

To examine whether SCOPE reflects embodied downstream behavior, we analyze correlations between SCOPE scores and performance on ManipBench (Zhao et al., 2025) and EmbodiedBench (Yang et al., 2025b).

SCOPE shows very high correlation with ManipBench (Pearson 0.88) and significant correlation with EmbodiedBench (Pearson 0.81), indicating that SCOPE captures viewpoint-dependent spatial subskills that align with downstream embodied performance and isolate a bottleneck underlying embodied failure.

C.6 Error Type Analysis

We analyze the error overlap between human participants and models on SCOPE-188 to assess whether model inconsistencies stem from perceptual ambiguity rather than reasoning failures. If single-view ambiguity caused model failures, human and model errors would be correlated, as both would struggle on the same ambiguous items (Table 9).

Model	Original	CoT	Gain
LLaMA3.2-11B-Vision	21.49	24.03	+2.54
LLaVA-OneVision-8B	24.88	26.90	+2.02
InternVL-78B	26.82	31.80	+4.98

Table 10: **Effect of CoT prompting on the Spatial Updating task.** CoT prompting consistently improves model performance across all evaluated models.

The near-zero correlations across all tasks ($|\text{Pearson}| \leq 0.25$) imply that model failures stem more from spatial reasoning limitations than from perceptual uncertainty inherent in single RGB images.

C.7 CoT Prompting Analysis.

To complement the single fixed prompting setup used in the main experiments, we additionally evaluate whether chain-of-thought (CoT) prompting (Kojima et al., 2022) (*"Let's think step by step,"* max tokens: 1024) can improve model performance on the Spatial Updating task. Prior work has shown that structured reasoning strategies can meaningfully improve performance on vision-language tasks (Kim et al., 2026; Kim and Cho, 2026). As shown in Table 10, CoT consistently raises accuracy across all three models, with gains ranging from **2.0** to **5.0** percentage points, and the performance gaps across models become more pronounced under CoT, suggesting that structured reasoning elicitation can enhance the diagnostic resolution of SCOPE. Nevertheless, all models remain well below human-level performance, indicating that the core spatial reasoning challenges captured by SCOPE persist regardless of prompting strategy. These results are consistent with our limitation discussion (Section 6), and we view broader exploration of prompting and decoding strategies as a promising avenue for future work.

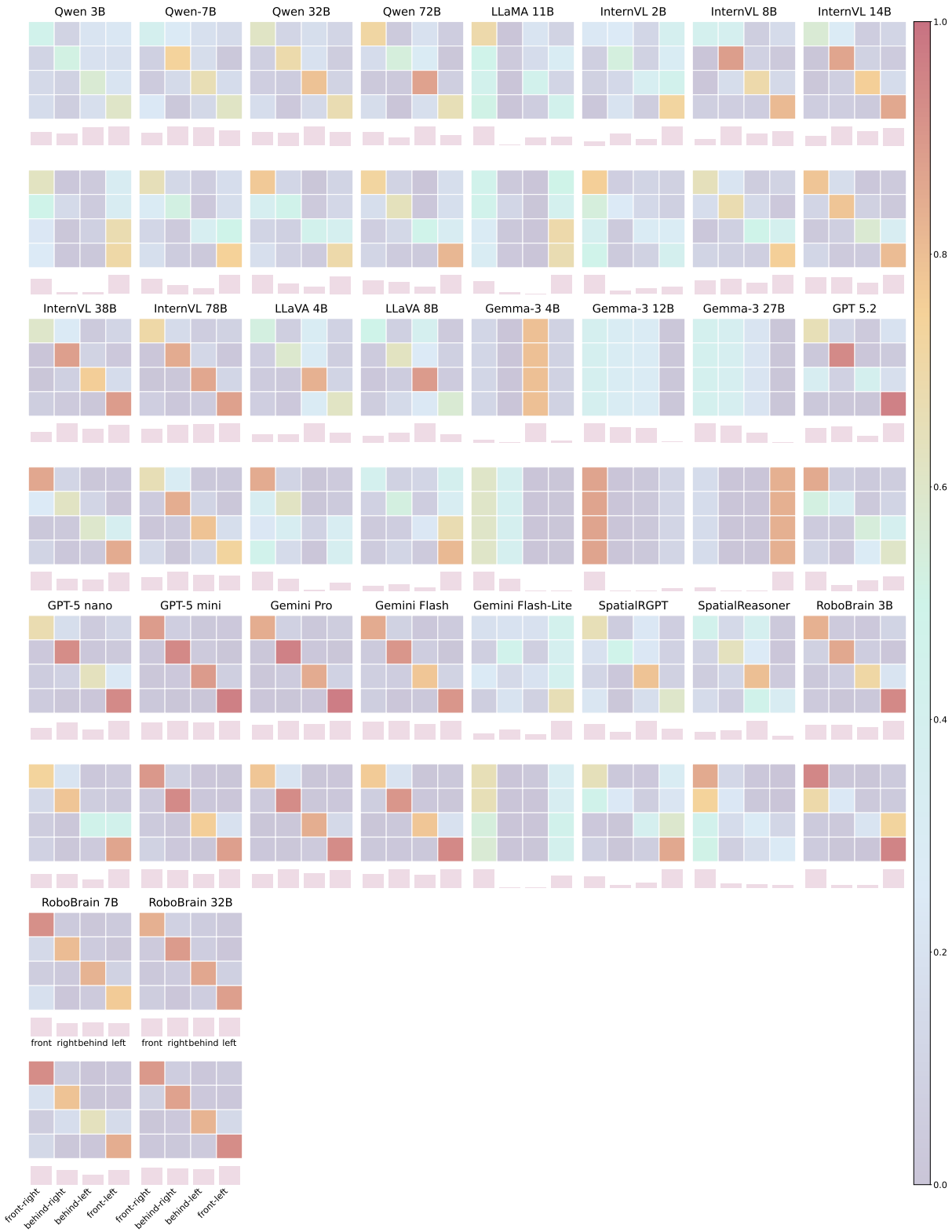
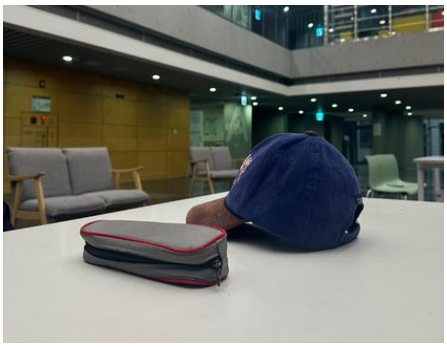


Figure 17: **Relational confusion matrices of all models for egocentric spatial reasoning.** For each pair of rows, the top row shows the orthogonal (front/right/behind/left) directions, while the bottom row shows the diagonal (front-left/front-right/behind-right/behind-left) directions.

Q) From the camera's perspective, where is the bench relative to the trach can?

A: front-right, B: front-left,
C: behind-right, D: behind-left

A) D



Q) From the camera's perspective, where is the cap relative to the pouch?

A: front-right, B: front-left,
C: behind-right, D: behind-left

A) C

Q) From the camera's perspective, where is the smartphone relative to the apple?

A: front, B: behind,
C: left, D: right

A) C

Q) From the camera's perspective, where is the green trash can relative to the orange trash can?

A: front-right, B: front-left,
C: behind-right, D: behind-left

A) B

Figure 18: **Spatial Consistency (Egocentric) examples.**

Q) From the perspective of a standing man, where is rock relative to a standing man?

A: front, B: behind, C: left, D: right

A) D

Q) Imagine you are sitting on the bench. From your perspective, where is the black tile relative to you?

A : front, B : behind, C: left, D: right

A) A

Q) From the perspective of a shepherd dog, where is bed relative to a shepherd dog?

A: front, B: behind, C: left, D: right

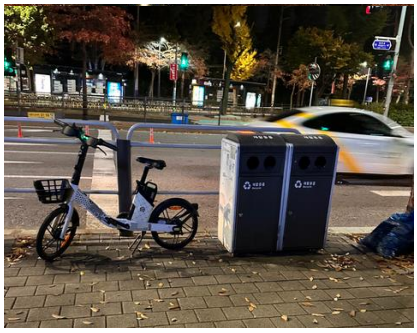
A) B

Q) Imagine you are sitting on the bench. From your perspective, where is metal drum relative to you?

A : front-right, B : behind-right, C: front-left, D: behind-left

A) D

Figure 19: **Spatial Consistency (Allocentric) examples.**



Q) Consider the scene from the viewer's perspective. Imagine a point exactly between the bicycle and the trash can. Rotate 90° counter-clockwise around this point. Now, where is the bicycle relative to the trash can?

A: front, B: behind,
C: right, D: left

A) B

Q) Consider the scene from the viewer's perspective. Imagine a point exactly between the plastic chair and the black parking sign. Rotate 90° counter-clockwise around this point. Now, where is the plastic chair relative to the black parking sign?

A: front-left, B: behind-right,
C: front-right, D: behind-left

A) C

Figure 20: **Spatial Updating examples. (Egocentric)**

Q) Imagine a point exactly between a standing man and a cat. Rotate 90° counter-clockwise around this point. Now, from the perspective of the man, where is the cat?

A: front, B: behind,
C: left, D: right

A) C

Q) Imagine a point exactly between the bench and the road. Rotate 180° counter-clockwise around this point. Now imagine you are sitting on the bench. From your perspective, where is the road relative to you?

A: front, B: behind,
C: left, D: right

A) A

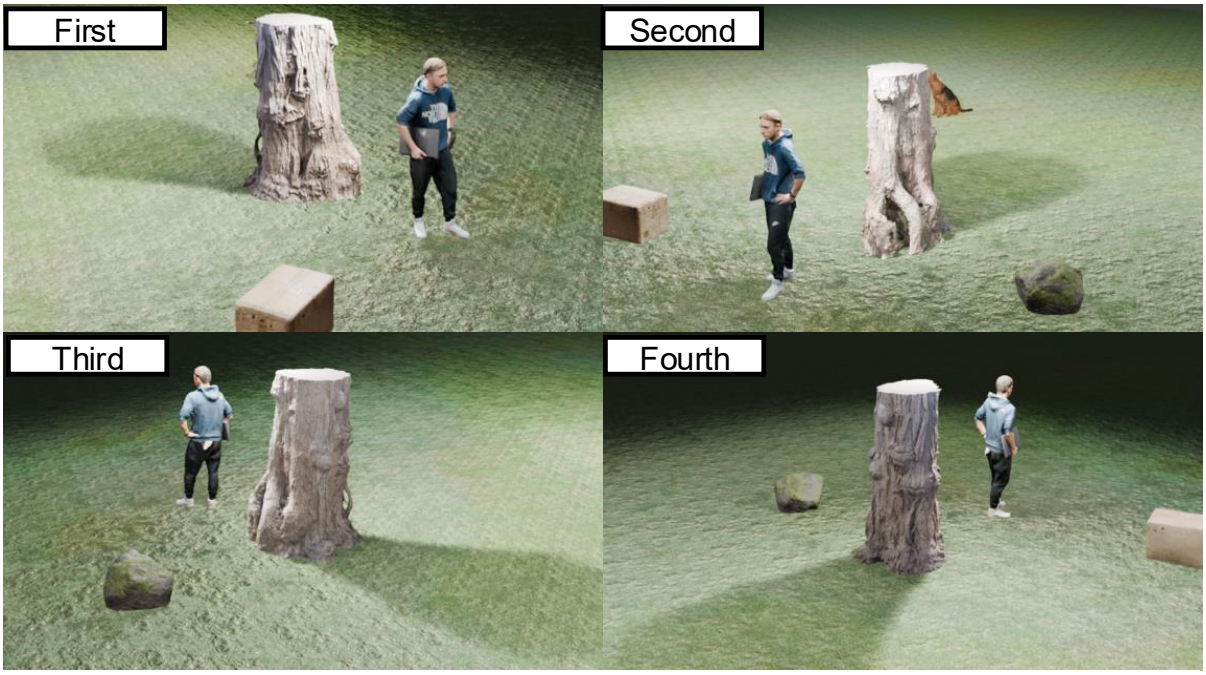
Figure 21: **Spatial Updating examples. (Allocentric)**

Q) From the camera's perspective, a cup is to the left of a black cellphone. When seeing these four images (First Image is 0°, Second Image is 90°, Third Image is 180°, Fourth Image is 270°), at which scene is this statement true? Answer with one of [First Image, Second Image, Third Image, Fourth Image].

A: First Image, B: Second Image,
C: Third Image, D: Fourth Image

A) A

Figure 22: **Spatial Integration examples. (Egocentric)**



Q) From the standing human's perspective, there is an object behind them. What is the nearest object to it? When seeing these four images (First Image is 45°, Second Image is 135°, Third Image is 225°, Fourth Image is 315°), answer with one of the given choices. Choices: [rock, bicycle, box, dog].

A: rock, B: bicycle,
C: box, D: dog

A) A

Figure 23: **Spatial Integration examples. (Allocentric)**