DM3SR: BENCHMARKING DEEP MULTI-VIEW 3D SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models show strong VQA performance but remain weak at multi-view 3D spatial reasoning. Many existing benchmarks, particularly those built from video-derived frames, provide limited coverage of viewpoint diversity and vertical scene structure, leaving model behavior under broader 3D configurations insufficiently characterized. We introduce **DM3SR** (**D**eep **M**ulti-View **3**D **S**patial **R**easoning), a geometry-controlled benchmark spanning three complementary scene families: CONTEXTUAL, CONTROLLED-SYN, and CONTROLLED-REAL. DM3SR defines twelve subtasks covering object counting, absolute and relative distance, and lateral and vertical positional reasoning. Evaluating 32 VLMs reveals substantial gaps from human performance, task-dependent benefits from scaling, and inconsistent transfer of spatial adaptation across scenes and axes. Controlled analyses indicate preferences for XY-dominated geometry and continuous viewpoints, alongside difficulty integrating local observations along dimensions with greater scene extent. DM3SR provides a testbed for evaluating cross-view integration and diagnosing how scene geometry and viewing conditions shape spatial reasoning.

1 INTRODUCTION

Vision–language models (VLMs) (Qwen Team, 2025; Wang et al., 2025b; OpenAI, 2026) have rapidly advanced on visual question answering benchmarks (Antol et al., 2015; Hudson & Manning, 2019; Johnson et al., 2017; Marino et al., 2019), yet still struggle with 3D spatial reasoning (Ma et al., 2024; Zhang et al., 2025b; Yang et al., 2024; Yu et al., 2025a). While early evaluations focused primarily on single-image tasks such as depth prediction and orientation estimation (Ma et al., 2024; Zhang et al., 2025b; Shiri et al., 2024), longer context windows have enabled multi-view (Yang et al., 2026; Yin et al., 2026; Chun-Hsiao Yeh & Ma, 2026) and video-based (Yang et al., 2024; Cheng et al., 2025; Liao et al., 2025) evaluation, where models are asked to infer scene structure, object relations, and viewpoint geometry from partial and occluded observations. Integrating such observations is central to physical-world intelligence, enabling embodied agents to navigate, manipulate, and interact in environments that cannot be fully observed from a single viewpoint (Yu et al., 2025a).

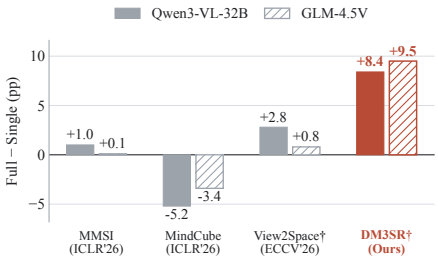


Figure 1: Full–Single accuracy gap (pp). DM3SR shows the largest multi-view gains for both models. † denotes the MCQ subset. Full results are in Appendix D.1.

However, many existing multi-view spatial reasoning benchmarks inherit viewing conditions and scene layouts from indoor scanning videos (Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021; Grauman et al., 2024; Ling et al., 2023; Xia et al., 2024). Continuous, eye-level camera trajectories limit elevation diversity and often yield redundant observations, while indoor object arrangements typically exhibit greater horizontal extent than vertical variation. These properties create two evaluation blind spots: **(1) spatial generalization**: reasoning beyond these scene and viewpoint distributions remains insufficiently tested; and **(2) cross-view integration**: individual views may contain sufficient evidence to answer multi-view questions, allowing success without integrating complementary information across views.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

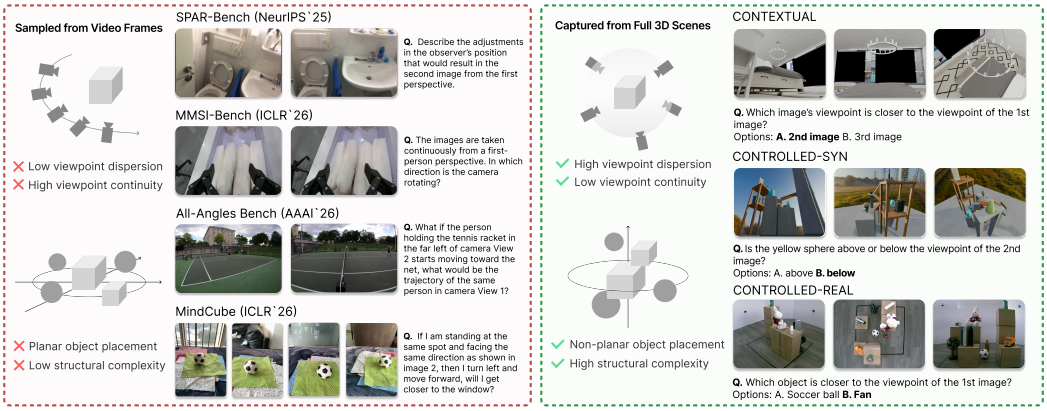


Figure 2: Comparison between DM3SR and existing multi-view spatial reasoning benchmarks (Zhang et al., 2025a; Yang et al., 2026; Chun-Hsiao Yeh & Ma, 2026; Yin et al., 2026). Prior benchmarks, often derived from indoor scanning videos, exhibit limited viewpoint diversity and planar object placement bias. In contrast, DM3SR provides diverse viewpoints over the full sphere and enhanced structural scene complexity.

To examine the benefit of multi-view inputs empirically, we compare two strong VLMs (Qwen Team, 2025; V Team et al., 2025) under full-view and single-view settings, where single-view accuracy is averaged over evaluations using each view separately. As shown in Fig. 1, the comparison benchmarks (Yang et al., 2026; Yin et al., 2026; Ke et al., 2026) exhibit small or negative gains from full-view inputs for these models. These findings highlight the need to distinguish benchmark accuracy from a model's ability to integrate complementary evidence across views.

Motivated by these limitations, we introduce **DM3SR** (**D**eep **M**ulti-View **3**D **S**patial **R**easoning), a benchmark designed to require and diagnose multi-view spatial reasoning under controlled variation in viewing conditions and scene geometry (Fig. 2). DM3SR renders or captures observations from complete 3D scenes and deliberately arranged physical scenes. It comprises three complementary scene families: CONTEXTUAL, CONTROLLED-SYN, and CONTROLLED-REAL. CONTEXTUAL provides semantically structured synthetic indoor environments, while the two CONTROLLED splits extend evaluation to object arrangements with greater vertical variation through synthetic generation and physical capture, respectively. Synthetic scene generation further enables targeted manipulation of geometric properties. Across these scene families, DM3SR defines twelve subtasks covering object counting, absolute and relative distance, and lateral and vertical relative positioning.

In the same comparison, DM3SR yields substantially larger full-view gains, indicating a greater empirical benefit from multi-view inputs. Evaluation of 32 open-source, proprietary, and spatially specialized VLMs reveals substantial gaps from human performance, task-dependent scaling trends, and inconsistent transfer of spatial adaptation across scene families and axes. Diagnostic experiments indicate preferences for continuous viewpoint sequences and XY-dominated geometry. Task-specific responses to additional views further suggest difficulty integrating local observations along dimensions in which scenes have greater spatial extent. Together, these findings highlight limitations in spatial reasoning across diverse scene geometries and viewing conditions.

Our contributions are as follows:

- We introduce DM3SR, a multi-view 3D spatial reasoning benchmark spanning CONTEXTUAL, CONTROLLED-SYN, and CONTROLLED-REAL scenes, with answers derived from scene geometry and controlled variation of viewpoints and object configurations.
- We evaluate 32 open-source, proprietary, and spatially specialized VLMs, characterizing substantial gaps from human performance, task-dependent benefits from model scaling, and variation in the gains from spatial adaptation across scene families and spatial axes.
- We conduct controlled analyses of view count, viewpoint order, and scene geometry, revealing preferences for continuous viewpoints and XY-dominated geometry, and suggesting difficulty integrating local observations along dimensions with greater scene extent.

2 Related Work

Spatial Reasoning Benchmarks.

Early benchmarks such as 3DSRBench (Ma et al., 2024), SPHERE (Zhang et al., 2025b), and Spatial-MM (Shiri et al., 2024) evaluate spatial reasoning from a *single* image. Video-based benchmarks (e.g., VSI-Bench (Yang et al., 2024)) extend evaluation to multi-frame inputs using indoor scanning datasets such as ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021), and introduce 2D cognitive maps for analyzing VLM behavior. More recent efforts, including Multi-SpatialMLLM (Xu et al., 2025), MMSI-Bench (Yang et al., 2026), SPAR-Bench (Zhang et al., 2025a), MindCube (Yin et al., 2026), and All-Angles Bench (Chun-Hsiao Yeh & Ma, 2026), sample multi-view images from videos (ScanNet, ARKitScenes, DL3DV-10K (Ling et al., 2023), Ego-Exo4D (Grauman et al., 2024)). Although these benchmarks move toward multi-view evaluation, they inherit the same trajectory-induced biases, restricting geometric complexity and occlusion diversity. Synthetic benchmarks such as Spatial457 (Wang et al., 2025d) and View2Space (Ke et al., 2026) instead use rendered scenes.

Spatial VLMs.

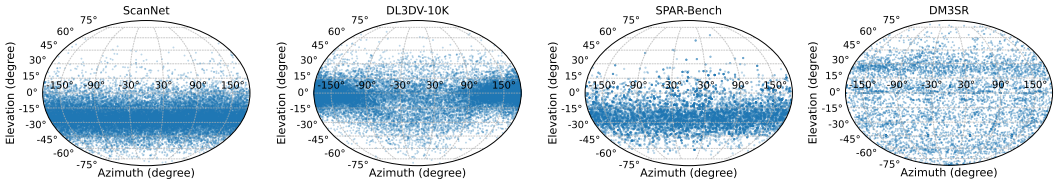
Single-image spatial VLMs such as SpatialVLM (Chen et al., 2024), SpatialRGPT (Cheng et al., 2024), ASMv2 (Wang et al., 2024), and SpaceQwen (Jia et al., 2025) aim to improve direction, depth, and geometric understanding from a single view. With longer context windows, models began to incorporate multi-view or video inputs. Approaches such as Multi-SpatialMLLM (Xu et al., 2025) and SPAR (Zhang et al., 2025a) perform supervised fine-tuning on multi-view datasets but remain limited by their training distributions. Reinforcement learning like GRPO (Shao et al., 2024) has recently improved spatial reasoning: SpaceR (Ouyang et al., 2025) and MindCube (Yin et al., 2026) employ answer rewards and 2D cogmap-based rewards on indoor videos, while ViLaSR (Wu et al., 2025) introduces tool-augmented RL with visual drawing supervision. VG-LLM (Zheng et al., 2025) improves spatial reasoning by incorporating VGGT (Wang et al., 2025a) as an auxiliary 3D geometry-aware encoder.

3 DM3SR

We introduce **DM3SR**, a benchmark explicitly designed to evaluate *multi-view 3D spatial reasoning* in vision–language models (VLMs). While existing multi-view benchmarks (Xu et al., 2025; Yang et al., 2026; Chun-Hsiao Yeh & Ma, 2026; Zhang et al., 2025a; Yin et al., 2026) have advanced the evaluation of spatial reasoning, many draw on *indoor scanning videos* (Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021; Grauman et al., 2024; Ling et al., 2023; Chang et al., 2017), typically captured along human-height camera trajectories. Such datasets exhibit three systemic biases that hinder the reliable measurement of multi-view reasoning:

- **Viewpoint bias:** viewpoints concentrate within a narrow elevation range (Fig. 3) and follow continuous trajectories, often yielding redundant views.
- **Object-placement bias:** object positions exhibit greater spread along the XY-plane than along the Z-axis, reflecting predominantly surface-bound layouts.
- **Structural simplicity:** regular indoor layouts provide strong semantic priors and limited structural diversity, while low object counts, clutter, and occlusion further simplify scenes.

Together, these biases limit the reliable evaluation of cross-view integration across diverse 3D configurations and spatial relations. To address these limitations, DM3SR renders diverse viewpoints from complete 3D scenes, enabling explicit control over object configurations, occlusion, and structural complexity for systematic evaluation of cross-view spatial integration. It comprises three complementary scene families, Contextual, Controlled-Syn, and Controlled-Real, which together provide contextual richness, geometric control, and real-world capture (Fig. 4).



(a) ScanNet. (b) DL3DV-10K. (c) SPAR-Bench. (d) DM3SR.

Figure 3: Viewpoint direction distributions across datasets shown in a Mollweide projection. The horizontal axis denotes azimuth and the vertical axis denotes elevation.

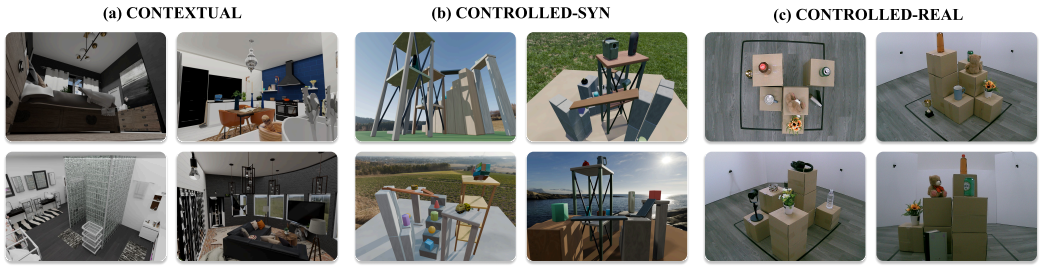


Figure 4: Overview of DM3SR scene families: (a) CONTEXTUAL synthetic indoor scenes with semantic context; (b) CONTROLLED-SYN, synthetic outdoor scenes with vertically varied object arrangements; (c) CONTROLLED-REAL, physically captured arrangements with controlled geometry.

3.1 SCENE FAMILIES AND GENERATION

3.1.1 CONTEXTUAL

This subset leverages realistic indoor layouts from the *Habitat Synthetic Scenes Dataset (HSSD)* (Khanna* et al., 2023). We select rooms larger than 20 m^2 and retain only regions where the number of objects per unit area exceeds a predefined threshold, thereby excluding sparsely populated spaces such as corridors. This yields semantically rich, cluttered scenes with realistic occlusions while retaining complete 3D ground truth.

3.1.2 CONTROLLED

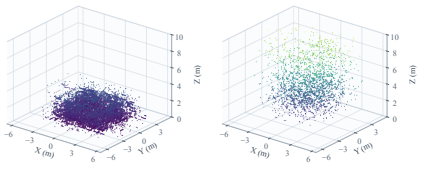
CONTROLLED-SYN provides precise geometric control for systematically isolating spatial factors. Scenes are rendered in *Blender* using 82 object categories from *OmniObject3D* (Wu et al., 2023) and six primitive shape categories, with 8–14 objects per scene. We introduce eight types of support structures whose positions, Z-axis rotations, dimensions, and numbers of tiers are randomly sampled to balance object-position variance across the X, Y, and Z axes. Physics-based validation filters out configurations with overlapping structures, unsupported objects, or ceiling contacts. Diverse outdoor HDRI[1] backgrounds are used to maintain photorealistic appearance.

For CONTROLLED-REAL, we construct tabletop scenes using stacked cardboard cubes (a 4×4 grid, 20 cm per cell) and arrange 6–9 everyday objects per scene, drawn from 28 objects across 15 categories. More construction details are in Appendix A.

3.2 QUANTITATIVE SPATIAL CONFIGURATION ANALYSIS

3.2.1 OBJECT DISTRIBUTION.

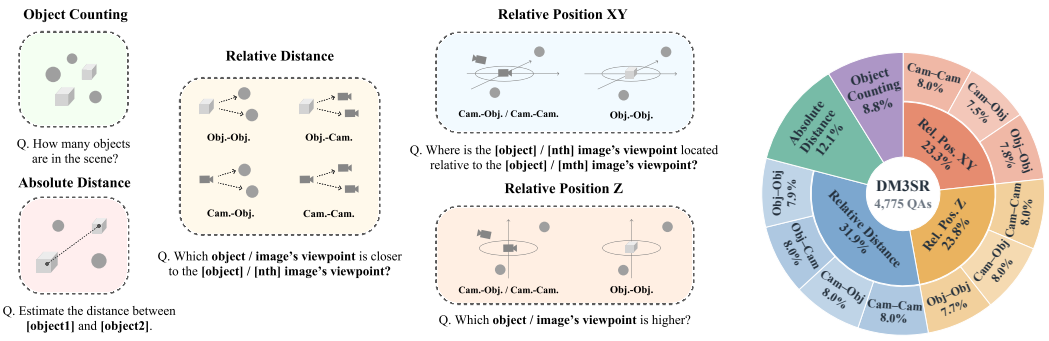
Conventional multi-view scenes exhibit flattened object-position distributions because objects lie primarily on the floor and the horizontal scene extent is substantially larger than its height. As shown in Fig. 5, this tendency is also evident in CONTEXTUAL. In contrast, CONTROLLED-



(a) CONTEXTUAL. (b) CONTROLLED.

Figure 5: Object position distributions.

[1]High Dynamic Range Imaging.



(a) Overview of benchmark tasks. (b) Task distribution.

Figure 6: Overview of DM3SR tasks. (a): Illustration of the five task families—Object Counting, Absolute Distance, Relative Distance, and Relative Position XY/Z—together with their twelve anchor–target configurations. (b): Proportion of QA pairs across task types.

SYN is constructed to yield comparable positional variance across the X, Y, and Z axes, enabling balanced evaluation of spatial reasoning along all three dimensions.

3.2.2 VIEWPOINT DIVERSITY.

All viewpoint directions are oriented toward the scene center. CONTROLLED-SYN uniformly samples 26 viewpoints over an enclosing sphere and the ground plane; CONTEXTUAL distributes viewpoints across the walls, floor, and ceiling in proportion to their surface areas; and CONTROLLED-REAL uses nine fixed viewpoints (four side, four diagonal, and one top-down). Viewpoint-direction diversity is quantified using two complementary metrics:

$$U_{\text{sph}} = 1 - \left\| \frac{1}{N} \sum_i \mathbf{v}_i \right\|, \quad U_{\text{ent}} = -\frac{\sum_{j=1}^{M} p_j \log p_j}{\log M} \tag{1}$$

where $\mathbf{v}_i$ denotes a unit viewing direction and p_j is the normalized frequency in the j-th bin of a 36×18 azimuth–elevation grid ($M = 648$). U_{sph} measures *directional balance*, i.e., how little the view directions are biased toward a dominant direction, while U_{ent} measures how widely the view directions spread across bins. Directions of all unique images are pooled per dataset. As summarized in Tab. 1 and Fig. 3d, DM3SR achieves the highest viewpoint diversity on both metrics, exceeding prior datasets such as ScanNet (Dai et al., 2017) and SPAR-Bench (Zhang et al., 2025a).

3.3 TASK FORMULATION AND QA GENERATION PIPELINE

3.3.1 TASK DESIGN.

As illustrated in Fig. 6, DM3SR evaluates five capabilities: (1) **Object Counting**, estimating the total number of unique objects across views; (2) **Absolute Distance**, estimating metric distances between object pairs; (3) **Relative Distance**, comparing distances among objects and camera viewpoints across four configurations; (4) **Relative Position XY**, reasoning about left/right and front/back relations; and (5) **Relative Position Z**, reasoning about higher/lower relations. The first two require numeric answers, while the remaining three use a multiple choice format, together comprising 12 subtasks. We use mean relative accuracy (MRA; see Appendix C.2.2) for Absolute Distance and accuracy for Object Counting and all multiple-choice tasks.

3.3.2 QA GENERATION AND QUALITY CONTROL.

For each scene, candidate objects and view subsets (2–6 views) are sampled until geometric constraints are satisfied. Task-specific templates then generate natural-language questions by inserting metadata such as category labels, view indices, and anchor types. In the CONTEXTUAL split, since many HSSD categories are incorrect or too coarse, every object in the target regions is relabeled with a specific class label generated by a VLM from four rendered views (Appendix A.3.1). Ambiguous or heavily occluded samples are discarded to preserve geometric validity.

Table 1: Comparison of viewpoint direction diversity across datasets.

Dataset	$\mathbf{U}_{\mathrm{sph}}$ (↑)	$\mathbf{U}_{\mathrm{ent}}$ (↑)
ScanNet	0.595	<u>0.819</u>
DL3DV-10K	0.534	0.740
SPAR-Bench	0.632	0.814
MMSI-Bench	0.371	0.598
MindCube	<u>0.675</u>	0.579
View2Space	0.609	0.677
DM3SR (Ours)	**0.905**	**0.923**

Table 2: Dataset statistics of DM3SR and existing multi-view spatial reasoning benchmarks. "Images" denotes the total number of rendered (for synthetic) or captured (for real) unique views.

Benchmark	QAs	Scenes	Images	Avg. Images per QA	Datasets
SPAR-Bench	7,207	-	-	2.04	ScanNet, ScanNet++
MMSI-Bench	1,000	-	1,990	2.55	ScanNet, Matterport3D
MindCube	21,154	975	3,268	3.37	ArkitScenes, DL3DV-10K
All-Angles	2,132	90	-	2.43	Ego-Exo4D, EgoHumans
CONTEXTUAL	1,883	149	1,838	3.92	HSSD
CONTROLLED-SYN	2,244	187	4,205	4.10	OmniObject3D
CONTROLLED-REAL	648	48	432	4.11	-
DM3SR (Total)	4,775	384	6,475	4.03	OmniObject3D, HSSD

In total, DM3SR contains **384 scenes**, **6,475 images**, and **4,775 QA pairs** across three splits. As summarized in Tab. 2, DM3SR provides substantially higher scene and image diversity per QA pair than existing benchmarks.

4 EVALUATION ON DM3SR

4.1 SETUP

We evaluate 32 VLMs on DM3SR, grouped into three categories. The closed source group consists of GPT-5.4 (OpenAI, 2026), Gemini-3.5-Flash (Google DeepMind, 2026), and Claude-Sonnet-5 (Anthropic, 2026). The open source group covers a broad range of capacities and architectures, including Qwen2.5/3-VL (Bai et al., 2025; Qwen Team, 2025), InternVL3.5 (Wang et al., 2025b), Ovis2.5 (Lu et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025b), Gemma 4 (Gemma Team, 2026) and GLM-4.5V (V Team et al., 2025). Finally, we include spatially adapted models that are explicitly tuned for spatial reasoning: SpaceR (Ouyang et al., 2025), ViLaSR (Wu et al., 2025), VG-LLM (Zheng et al., 2025), MindCube (Yin et al., 2026), SpaceQwen, and SpaceThinker (Jia et al., 2025).

In addition to model results, we report three reference baselines. The *Random Guess* baseline corresponds to uniform guessing, and the *Majority Option* baseline reflects majority-class accuracy under each task's label distribution. The *Human* baseline is measured on a 240-QA subset of DM3SR (DM3SR-TINY), where annotators answer questions using the same visual inputs as the models.

4.2 MAIN RESULTS

4.2.1 OVERALL PERFORMANCE.

Tab. 3 summarizes the evaluation results across all tasks and scene splits. A large gap remains between human performance and all current VLMs. Among proprietary models, Gemini-3.5-Flash consistently achieves the best results across most tasks and scene types. Within the open-source group, Gemma 4-31B and Qwen3-VL-32B attains the highest overall accuracy and performs competitively with GPT-5.4. These results show that large open-source models can approach proprietary systems on DM3SR, yet all models still fall far short of human performance. Full results for all evaluated models are provided in Appendix C.3.

4.2.2 SCENE DISTRIBUTION.

As described in Sec. 3.1, CONTEXTUAL retains typical indoor semantics and predominantly XY-oriented object layouts, whereas CONTROLLED has more balanced positional variance across all three axes and more diverse 3D geometric relations. Relative to their Qwen2.5-VL-7B base, SpaceR, ViLaSR, and VG-LLM-8B show a mean relative accuracy change of $+5.6\%$ on CONTEXTUAL but -4.6% on CONTROLLED (Tab. 3). Across the five models in Fig. 7, Relative Position XY improves by 25.3% on average on CONTEXTUAL, while Relative Position Z declines by 8.8% on CONTROLLED. These results suggest that spatial adaptation favors XY-oriented reasoning in famil-

Table 3: DM3SR evaluation results across models. Performance gains over each model's base variant are also reported for the spatial models. Green cells denote the top two models in *overall* performance, while gray cells mark the top two scores within each task and split (dark = 1st, light = 2nd).

Model	Overall	By Task					By Scene Split		
	Avg.	Obj. Cnt.	Abs. Dist.	Rel. Dist.	Rel. Pos. XY	Rel. Pos. Z	CONTE-XTUAL	CONTROLLED SYN	CONTROLLED REAL
Random Guess	44.5	–	–	50.0	40.3	41.1	41.0	46.7	46.5
Majority Option	47.0	–	–	51.1	47.8	40.7	48.7	45.0	49.8
DM3SR-tiny									
Human	79.0	75.0	61.5	91.7	83.3	83.3	66.9	80.0	90.0
GPT-5.4 (OpenAI, 2026)	54.4	47.9	42.9	60.4	41.7	79.2	38.8	61.3	63.3
Qwen3-VL-32B (Qwen Team, 2025)	50.5	35.4	44.2	66.7	47.9	58.3	38.4	48.8	64.4
Proprietary Models (API)									
GPT-5.4 (OpenAI, 2026)	56.8	31.0	41.6	61.5	46.4	77.9	52.0	59.0	63.1
Gemini-3.5-Flash (Google DeepMind, 2026)	61.7	41.5	52.1	64.4	46.3	85.6	56.4	64.9	66.2
Claude-Sonnet-5 (Anthropic, 2026)	55.7	20.9	46.4	58.1	51.0	74.7	52.9	54.8	67.1
Open-source Models									
Qwen2.5-VL-7B (Bai et al., 2025)	47.6	20.1	38.6	54.2	44.1	56.9	44.9	48.3	53.1
Qwen2.5-VL-72B (Bai et al., 2025)	48.6	17.1	39.7	55.4	40.3	63.8	43.7	50.0	57.8
Qwen3-VL-8B (Qwen Team, 2025)	46.5	24.4	26.5	59.5	39.8	54.1	40.7	51.0	47.9
Qwen3-VL-32B (Qwen Team, 2025)	55.8	34.4	37.5	64.5	50.3	66.7	49.3	59.6	61.4
InternVL3.5-8B (Wang et al., 2025b)	46.9	22.5	43.4	53.9	38.7	56.2	44.4	47.1	53.1
InternVL3.5-38B (Wang et al., 2025b)	49.3	22.5	37.7	58.5	43.1	58.6	45.8	50.6	54.7
Ovis2.5-9B (Lu et al., 2025)	47.4	6.6	44.1	56.7	41.9	57.3	48.1	43.7	58.3
MiniCPM-V-4.5 (Yu et al., 2025b)	42.8	13.3	36.7	45.4	42.7	53.4	40.5	43.3	47.7
GLM-4.5V (V Team et al., 2025)	54.3	19.4	48.1	63.1	46.4	66.4	50.8	54.6	63.5
Gemma-4-12B (Gemma Team, 2026)	48.2	14.7	37.1	53.8	42.5	64.2	46.0	49.2	51.0
Gemma-4-31B (Gemma Team, 2026)	57.2	38.6	46.8	60.5	45.6	76.3	51.6	59.4	65.9
Spatial Models									
SpaceR (Ouyang et al., 2025)	47.8 (+0.2)	18.5	36.2	54.1	46.2 (+2.1)	57.7 (+0.8)	46.6 (+1.8)	47.5 (-0.8)	52.3 (-0.8)
ViLaSR (Wu et al., 2025)	48.2 (+0.6)	19.9	37.9	54.3	47.3 (+3.2)	56.7 (-0.3)	47.6 (+2.8)	47.3 (-1.0)	53.1 (+0.0)
VG-LLM-8B (Zheng et al., 2025)	45.6 (-2.0)	8.3	29.7	56.5	44.7 (+0.5)	53.8 (-3.2)	47.8 (+3.0)	45.0 (-3.3)	41.2 (-11.9)
MindCube-RawQA (Yin et al., 2026)	42.3 (-1.0)	0.2	40.0	50.6	47.6 (+5.8)	42.7 (-4.7)	43.6 (-0.7)	39.9 (-0.9)	46.6 (-1.9)
MindCube-CogMap (Yin et al., 2026)	38.5 (-4.8)	3.3	19.3	49.2	45.6 (+3.9)	39.8 (-7.6)	37.3 (-7.1)	39.1 (-1.7)	39.6 (-9.0)
SpaceQwen2.5-VL-3B (Jia et al., 2025)	34.2 (-9.1)	6.2	23.7	47.9	22.6 (-19.2)	43.0 (-4.4)	31.8 (-12.6)	34.5 (-6.3)	40.1 (-8.5)

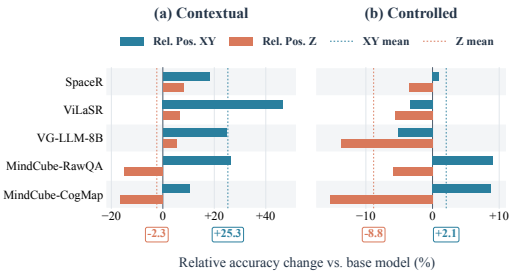


Figure 7: Relative accuracy changes (%) from spatial adaptation over base models across CONTEXTUAL and CONTROLLED scenes.

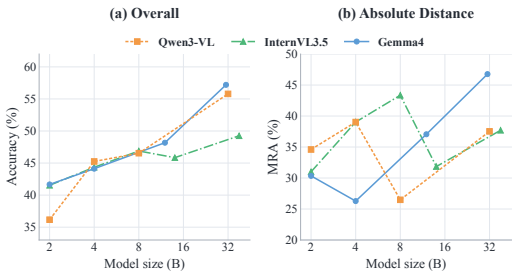


Figure 8: Comparison of overall and Absolute Distance task accuracy with respect to model size.

iar indoor settings, but its benefits do not consistently extend to other scene distributions or axes, with Z-axis reasoning on CONTROLLED showing degradation.

4.2.3 SCALING WITH MODEL SIZE.

Fig. 8 compares overall and Absolute Distance performance across model sizes in the Qwen3-VL, InternVL3.5, and Gemma4 families. While overall accuracy generally increases with model size, Absolute Distance fluctuates substantially across the Qwen3-VL and InternVL3.5 families, whereas Gemma4 shows a clearer improvement with scale. Relative Position XY also exhibits non-monotonic behavior among models below 14B parameters. In contrast, Object Counting, Relative Distance, and Relative Position Z scale more consistently. Overall, increased model capacity generally improves performance, but its effect varies considerably across spatial reasoning tasks and model families.

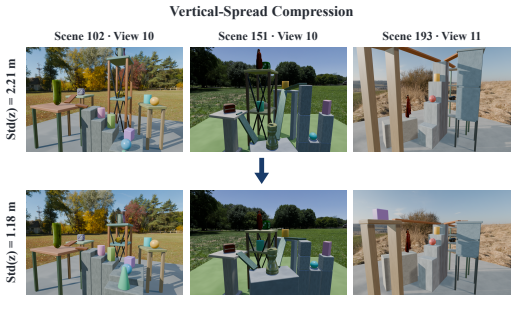


Figure 9: Paired scenes before (top) and after (bottom) vertical-spread compression.

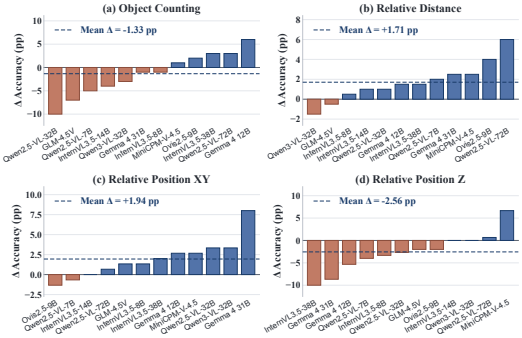


Figure 10: Task-wise accuracy changes (pp) across models after vertical-spread compression.

5 Diagnostic Analysis of Spatial Reasoning

Sec. 4.2.2 shows that gains from spatial adaptation vary across scene distributions and tasks. This raises the question of whether modern VLMs exhibit shared sensitivities to viewing conditions and scene geometry, potentially shaped by their training data and architectures. We use DM3SR to examine these sensitivities through controlled diagnostic experiments addressing three questions: (1) How does performance change with the spatial distribution of objects? (Sec. 5.1) (2) How does viewpoint continuity affect performance? (Sec. 5.2) (3) How effectively do VLMs use additional input views? (Sec. 5.3)

5.1 Spatial Layouts

We examine how the vertical object distribution affects spatial reasoning by reducing the normalized vertical spread from 1.0 to 0.5 in 50 paired Controlled-Syn scenes. As shown in Fig. 9, this reduces the standard deviation of object heights from 2.21 m to 1.18 m. Object identities, horizontal positions, camera views, and questions are held fixed, while answers are recomputed from the modified geometry. We evaluate 12 VLMs on 600 paired questions.

Fig. 10 shows that vertical compression improves Relative Distance and Relative Position XY accuracy for most models, by 1.71 and 1.94 pp on average, respectively, while reducing Relative Position Z accuracy by 2.56 pp. The Z decline is expected as height differences become less distinguishable. More notably, XY reasoning improves despite unchanged horizontal positions, and distance estimation benefits from reduced vertical variation. These gains suggest that current VLMs reason more effectively when vertical cues contribute less, revealing a preference for XY-dominated geometry. This pattern aligns with Sec. 4.2.2 and may reflect training data emphasizing horizontal arrangements and relations. Such a bias could limit generalization to tasks requiring spatial reasoning across all three axes, such as robotic manipulation and aerial navigation, underscoring the need for training and evaluation with more diverse spatial configurations.

5.2 Viewpoint Continuity

To quantify the continuity of camera viewpoints, we measure the *Mean Angular Deviation (MAD)* between consecutive camera directions. Let each input view be represented by a normalized camera direction vector $\mathbf{v}_i \in \mathbb{R}^3$ for $i = 1, \ldots, N$. MAD is defined as:

$$\text{MAD} = \frac{1}{N-1} \sum_{i=1}^{N-1} \arccos\left(\mathbf{v}_i \cdot \mathbf{v}_{i+1}\right) \tag{2}$$

A smaller MAD indicates smoother viewpoint transitions.

To construct the continuity-controlled groups, we exhaustively search all orderings of each question's views (at most six) and select the MAD-minimizing one, reducing the total angular path length by 30.0% (Appendix D.2.2). We evaluate eleven open-source VLMs under these reordered-view settings.

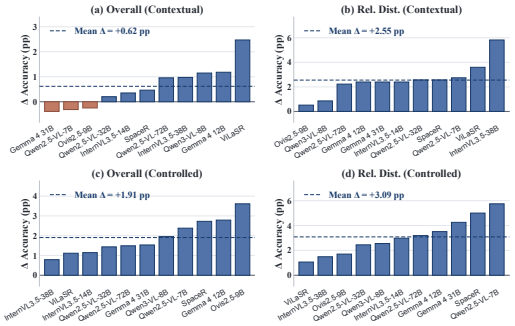


Figure 11: Accuracy changes (pp) from viewpoint-continuity control, comparing reordered and original view sequences.

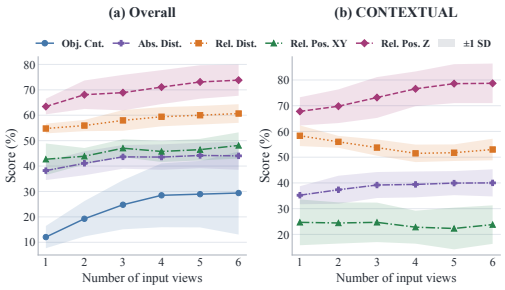


Figure 12: Performance variation across tasks with respect to the number of input images, averaged across the six open-source models.

Fig. 11 shows that continuity control improves aggregate performance for every evaluated model. Gains are consistent on CONTROLLED, averaging 1.91 pp, whereas several models degrade on CONTEXTUAL. Relative Distance benefits most consistently, with every model improving on both distributions by 2.55 and 3.09 pp on average, respectively. Since reordering preserves the visual evidence, these gains reveal sensitivity to viewpoint order, potentially reflecting familiarity acquired through video-based training. Benchmarks that retain the temporal order of source videos may therefore favor an input structure familiar to these models and overstate the robustness of cross-view integration. Success on continuous sequences does not establish the ability to integrate discontinuous views, as may be required in real-world settings involving spatially separated cameras or independently captured images.

5.3 MULTI-VIEW CAPACITY

We examine whether VLMs can effectively exploit additional views and how this ability varies across tasks and scene distributions. We sample 400 QA pairs from each scene split, yielding 1,200 pairs, and vary the input view count from one to six by randomly adding or removing one view at a time. We evaluate six strong open-source VLMs with at least 30B parameters. Fig. 12 reports their mean task scores, with shaded regions indicating ± 1 standard deviation across models. Full results are provided in Appendix D.2.1.

When averaged across splits, all tasks generally improve with more views, indicating that these models can benefit from additional visual evidence. However, CONTEXTUAL reveals a clear task-specific contrast: Relative Position Z improves with more views, whereas Absolute Distance largely saturates after three views, Relative Position XY shows little net gain, and Relative Distance declines from the first additional view. Compared with the compact CONTROLLED scenes, CONTEXTUAL scenes span a substantially wider extent along the XY plane, without a comparable expansion along Z. Integrating views of distinct local regions across this wider horizontal extent may increase the correspondence burden, making additional observations distracting rather than complementary. This axis-dependent pattern suggests that current VLMs struggle to integrate spatial information from local observations along dimensions in which the scene extends over a larger range.

6 CONCLUSION

We introduce DM3SR to address scene and viewpoint biases that existing benchmarks inherit from their source datasets. Built through direct rendering or capture of three complementary families of complete 3D scenes, DM3SR combines natural semantic context, diverse spatial relations, and geometric control across 12 subtasks. Evaluation of 32 VLMs reveals substantial gaps from human performance, task-dependent scaling trends, and uneven benefits from spatial adaptation across scene distributions. Diagnostic experiments further reveal preferences for XY-dominated geometry and continuous viewpoints, alongside difficulty integrating local observations along dimensions with greater scene extent. By enabling controlled variation of scenes and viewpoints, DM3SR supports both the evaluation of cross-view integration and targeted diagnosis of the limitations that must be addressed for more general 3D spatial reasoning.

AI USE STATEMENT

We used generative AI in three parts of this work. (1) *Dataset construction:* a vision–language model generated specific object class labels for the CONTEXTUAL split from four rendered views of each object, using the original HSSD category as a prior (Appendix A), and flagged CONTROLLED-SYN views blocked by nearby geometry. The generated labels only name objects in question text; all ground-truth answers are computed from 3D scene annotations. (2) *Coding:* an AI coding assistant helped implement and debug data-processing, evaluation, and plotting scripts. (3) *Writing:* AI assisted in drafting the appendix and revising the main text for clarity. The authors ran and verified all AI-assisted code against the underlying experiment records, inspected samples of the generated labels against the renders, and reviewed and edited all AI-assisted text. We take full responsibility for the content of this work.

ETHICS STATEMENT

DM3SR is built from HSSD (Khanna* et al., 2023) (CC BY-NC 4.0), OmniObject3D (Wu et al., 2023) (CC BY 4.0), Poly Haven HDRIs (CC0), and author-captured photographs of everyday objects; the images contain no people or personally identifiable information. The human evaluation interface records only a pseudonymous ID and the answers.

REPRODUCIBILITY STATEMENT

Scene construction is described in Sec. 3 and Appendix A, question templates in Appendix B, evaluation prompts, scoring, API identifiers, and full results in Appendix C, and the diagnostic protocols in Appendix D. We will release the benchmark data, 3D scene specifications, and generation and evaluation code.

REFERENCES

Anthropic. Claude Sonnet 5 system card. Technical report, Anthropic, June 2026. URL https://www.anthropic.com/claude-sonnet-5-system-card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

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

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning, 2025. URL https://arxiv.org/abs/2503.11495.

Shengbang Tong Ta-Ying Cheng Rouyu Wang Tianzhe Chu Yuexiang Zhai Yubei Chen Shenghua Gao Chun-Hsiao Yeh, Chenyu Wang and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *AAAI*, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Google DeepMind. Gemini 3.5 Flash model card. Technical report, Google DeepMind, May 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Fucai Ke, Zhixi Cai, Boying Li, Long Chen, Beibei Lin, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, and Hamid Rezatofighi. VIEW2SPACE: Studying multi-view visual reasoning from sparse observations. In *European Conference on Computer Vision (ECCV)*, 2026.

Mukul Khanna*, Yongsen Mao*, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic

Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for Object-Goal Navigation. *arXiv preprint*, 2023.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *ICCV*, 2023.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision, 2023. URL https://arxiv.org/abs/2312.16256.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report, 2025. URL https://arxiv.org/abs/2508.11737.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

OpenAI. GPT-5.4 Thinking system card. Technical report, OpenAI, March 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=b550154aa5ba6b812cdebba2b9dc1156c4369d40&from=research.research-list, 2025. Accessed: 2025-11-11.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, and Xiangyu Yue. Spacevista: All-scale visual spatial reasoning from mm to km, 2025. URL https://arxiv.org/abs/2510.09606.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation, 2025c. URL https://arxiv.org/abs/2505.05456.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M. de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025d.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025. URL https://arxiv.org/abs/2506.09965.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. In *ICLR*, 2026.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are vlms from true visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025a.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025b. URL https://arxiv.org/abs/2509.18154.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yujie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *NeurIPS*, 2025a.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling Spatial Blind Spots in Vision-Language Models Through Hierarchical Evaluation. *ACL*, 2025b.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *NeurIPS*, 2025.

Appendix

CONTENTS

A Dataset Construction **16**

A.1 Data Source Comparison 16

A.2 CONTROLLED-SYN 16

A.3 CONTEXTUAL 16

A.4 CONTROLLED-REAL 16

B QA Generation Pipeline **18**

B.1 Question Templates 18

B.2 QA Examples 18

C Evaluation on DM3SR **18**

C.1 Model APIs 18

C.2 Implementation Details 18

C.3 Complete Evaluation Results 20

D Diagnostic Analysis of Spatial Reasoning **21**

D.1 Motivation: Full-View versus Single-View Evaluation 21

D.2 Control Group Generation Details 22

D.3 Full Experiment Results 23

E Licenses and Limitations **24**

E.1 Licenses 24

E.2 Limitations 24

A Dataset Construction

A.1 Data Source Comparison

A.1.1 Data Sources of Spatial Benchmarks.

Tab. 4 lists the source datasets used by representative spatial reasoning benchmarks. Several reuse indoor scans or human-captured videos, whose camera trajectories can limit viewpoint coverage. Individual benchmarks may apply additional sampling or filtering.

Table 4: Source datasets used in spatial reasoning benchmarks.

Benchmark	Datasets
VSI-Bench (Yang et al., 2024)	ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021).
MMSI-Bench (Yang et al., 2026)	ScanNet (Dai et al., 2017), Matterport3D (Chang et al., 2017)
MultiSPA (Xu et al., 2025)	ScanNet (Dai et al., 2017).
SPAR-Bench (Zhang et al., 2025a)	ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023).
MindCube (Yin et al., 2026)	ArkitScenes (Baruch et al., 2021), DL3DV-10K (Ling et al., 2023).
All-Angles Bench (Chun-Hsiao Yeh & Ma, 2026)	Ego-Exo4D (Grauman et al., 2024), EgoHumans (Khirodkar et al., 2023).
SpaceVista (Sun et al., 2025)	ScanNet (Dai et al., 2017), DL3DV-10K (Ling et al., 2023).
SITE (Wang et al., 2025c)	Ego-Exo4D (Grauman et al., 2024).

A.2 Controlled-Syn

A.2.1 Scene Generation Pipeline.

Each retained Controlled-Syn scene contains 8–14 objects drawn from OmniObject3D scans (Wu et al., 2023) and six primitive shape categories. The available scans span 82 object categories. The current generator places objects on procedurally generated support structures. It samples eight support-structure types and varies their positions, Z-axis rotations, dimensions, and tier counts to distribute object centers across all three axes. Tab. 6 summarizes the scene components. Fig. 13 illustrates all eight support-structure types, all six primitive shapes, and representative OmniObject3D scans used by the generator.

We settle the objects with rigid-body physics and reject scenes with overlapping structures, unsupported objects, or ceiling contacts. For each retained scene, Blender renders 26 RGB views against an outdoor HDRI background. The retained set contains 187 scenes and 2,244 QA pairs. The 4,205 images reported for this split in Tab. 2 are the distinct rendered views referenced by its QA pairs; the generation protocol renders 26 views for each retained scene.

A.3 Contextual

A.3.1 Category Label Refinement.

Contextual uses HSSD (Khanna* et al., 2023) rooms larger than 20 m^2, restricted to regions above the object-density threshold in Sec. 3.1. Many HSSD categories are incorrect or too coarse to identify an object unambiguously (e.g., a half-moon console table labeled *dresser*, an egg chair labeled *seat*). We therefore relabel all 3,992 objects in the target regions. Each object is rendered in isolation from four elevated diagonal views (azimuths $45°/135°/225°/315°$, elevation $32°$), and Qwen3.8-27B receives these renders with the original category as a fallible prior. With greedy, JSON-constrained decoding, it outputs a specific class label of at most seven words (94% contain three to five) naming the object type and its distinguishing color or shape. These labels replace the original categories in all Contextual questions and options; answers, images, and geometry are unchanged. Fig. 14 and Tab. 7 show examples.

A.4 Controlled-Real

A.4.1 Physical Setups.

Fig. 15 shows the physical setup used to construct the Controlled-Real scenes in a 5.8 $\times$ 2.9 $\times$ 2.6 m room. We arrange stacked cardboard cubes on a 4×4 grid with 20 cm cells, surround the



Figure 13: Scene components used in CONTROLLED-SYN. Top: all eight support-structure types. Middle: all six primitive shape categories. Bottom two rows: twelve examples from the 82 OmniObject3D categories. Each component is rendered in isolation on a neutral studio floor using the benchmark's actual geometry and materials; apparent sizes are normalized for readability.

setup with white foam boards, and capture it from nine fixed viewpoints: four side, four diagonal, and one top-down. As summarized in Tab. 5, the object pool contains 28 objects from 15 categories. Each scene contains 6–9 objects; 48 scenes yield 432 captured views and 648 QA pairs.

Table 5: Object categories in the CONTROLLED-REAL split.

Category	# of objects
Book	2
Can	2
Digital clock	1
Fan	1
Headset	1
Mug	3
Paper cup	2
Party hat	1
Pet bottle	2
Plant	2
Plush toy	3
Scotch tape	1
Sports ball	3
Toy car	3
Trophy	1
Total	28

Table 6: CONTROLLED-SYN scene components.

Category	Components
Scanned objects	82 OmniObject3D object categories.
Primitive objects	Six shapes: box, sphere, cylinder, cone, hexagonal prism, and icosphere.
Support structures	Eight types: table, scaffold, arch, stairs, stacked blocks, single block, pedestal, and shelf. Their positions, orientations, dimensions, and tier counts vary.

B QA GENERATION PIPELINE

B.1 QUESTION TEMPLATES

Tab. 8 gives representative templates for the five task families and twelve anchor–target configurations in Sec. 3.3. Braced placeholders are filled from scene metadata. Object Counting uses split-specific wording: in CONTROLLED-SYN, questions distinguish plain primitives from scanned objects and exclude the support structures; in CONTROLLED-REAL, the cardboard supports are excluded. Absolute Distance requests a number in the stated unit and range. Relative Distance and Relative Position use multiple-choice options.

B.2 QA EXAMPLES

Figs. 16–18 show example questions from the three scene families.

C EVALUATION ON DM3SR

C.1 MODEL APIS

Tab. 9 records the API identifiers as of September 19, 2026. The model IDs and version conventions follow the provider documentation for GPT-5.4, Claude, and Gemini.

C.2 IMPLEMENTATION DETAILS

C.2.1 EVALUATION PROMPTS.

The default evaluation prompts distinguish numerical tasks (Object Counting and Absolute Distance) from multiple-choice tasks (Relative Distance and Relative Position XY/Z). They ask for a single number or option letter inside `<answer>...</answer>` tags. Model-specific variants are used where required by an adapted model's evaluation procedure. Figs. 19–22 show the default and MindCube-RawQA prompt formats.

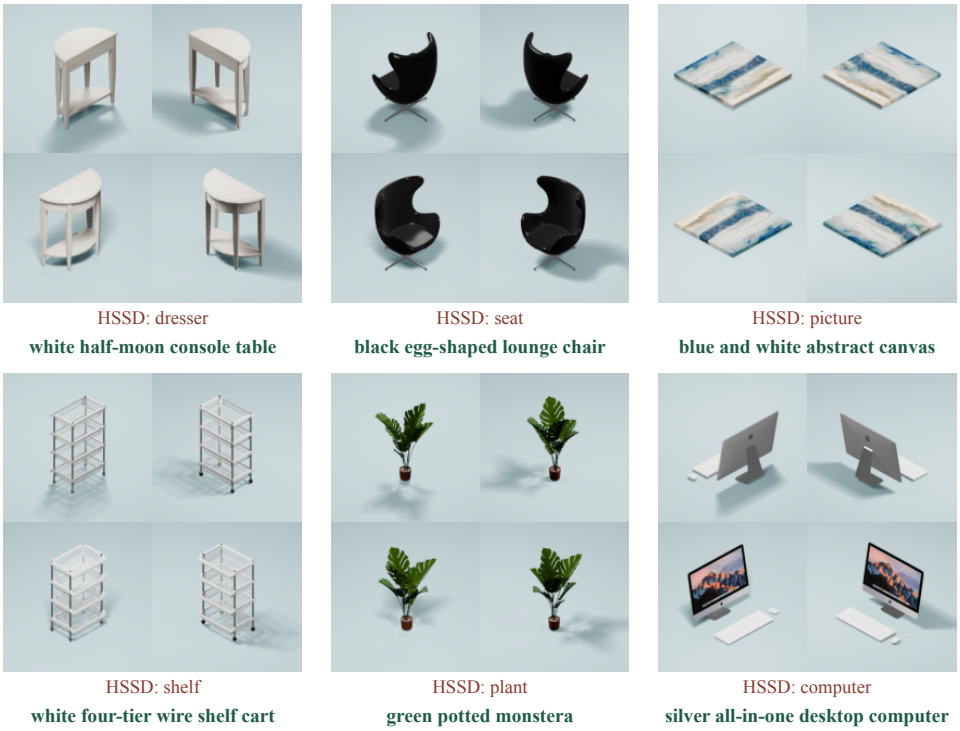


Figure 14: CONTEXTUAL relabeling examples. Each object is rendered from four elevated diagonal viewpoints and labeled by Qwen3.8-27B. Red: original HSSD category; green: generated class label used in the questions.

Table 7: Additional CONTEXTUAL relabeling examples: original HSSD category, generated broad family, and generated class label used in the questions.

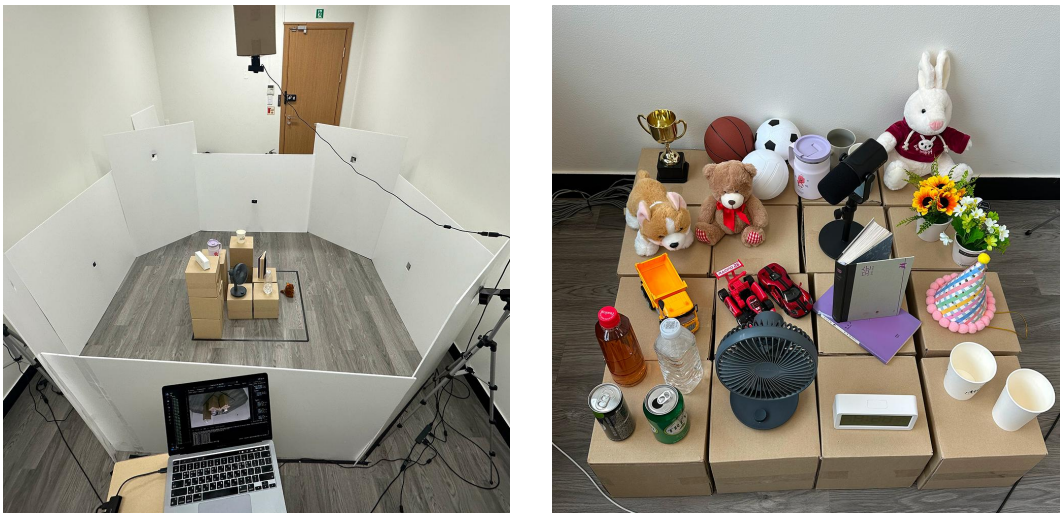
HSSD category	Broad family	Generated class label
lamp	lighting	gray circular recessed ceiling light
table	table furniture	round gray top wooden cross-leg side table
cabinet	storage furniture	white two-door cabinet with black top
book	book	maroon hardcover book
couch	seating furniture	white two-seat sofa
pillow	decorative cushion	green square throw pillow
vase	decorative vessel	dark glass bottle with rope net

C.2.2 SCORING.

Object Counting and the multiple-choice tasks use exact-match accuracy. Absolute Distance uses mean relative accuracy (MRA) (Yang et al., 2024), which, like the threshold-averaged metrics of PASCAL VOC (Everingham et al., 2010) and COCO (Lin et al., 2014), averages accuracy over multiple tolerance thresholds; we use ten relative-error thresholds. For a valid numerical prediction $\hat{d}$ and ground truth $d > 0$, the per-question score is

$$s_{\mathrm{MRA}}(\hat{d}, d) = \frac{1}{10} \sum_{k=1}^{10} \mathbb{1}\left[\frac{|\hat{d} - d|}{d} < 0.05k \right]. \tag{3}$$

The Absolute Distance score averages $100 s_{\mathrm{MRA}}$ over its 578 questions; an absent or invalid numeric answer receives zero. The overall score is the question-weighted mean over all 4,775 questions. Random Guess and Majority Option are defined for multiple-choice questions only.



(a) Camera and object settings. (b) Objects overview.

Figure 15: Physical setup and object pool for CONTROLLED-REAL scenes.

Table 8: Representative question templates. Braced terms denote scene-specific values.

Task	Subtask	Question Template
Object Counting	–	CONTROLLED-SYN: How many plain untextured (or photo-realistic scanned) objects are placed on the support structures? Exclude the other object type and the structures. CONTROLLED-REAL: How many objects are in the scene, excluding the cardboard box and background?
Absolute Distance	Obj.-Obj.	Estimate the 3D straight-line distance in { *unit* } between the centers of { *object1* } and { *object2* }. The answer must be between { *lower* } and { *upper* }.
Relative Distance	Cam.-Cam.	Which image's viewpoint is closer to the viewpoint of the { *nth* } image?
	Cam.-Obj.	Which object is closer to the viewpoint of the { *nth* } image?
	Obj.-Cam.	Which image's viewpoint is closer to the { *object* }?
	Obj.-Obj.	Which object is closer to the { *object* }?
Relative Position XY	Cam.-Cam.	Standing at the { *nth* } image's viewpoint and facing the scene center, where is the { *mth* } image's viewpoint?
	Cam.-Obj.	Relative to the viewing direction of the { *nth* } image, where is the { *object* }?
	Obj.-Obj.	Suppose you are standing at { *object1* } and facing { *object2* }. Where is { *object3* }?
Relative Position Z	Cam.-Cam.	Which image's viewpoint is the highest or lowest?
	Cam.-Obj.	Is { *object* } above or below the viewpoint of the { *nth* } image?
	Obj.-Obj.	Which object's center is the highest or lowest?

C.2.3 HUMAN EVALUATION INTERFACE.

Fig. 23 shows the interface used for the DM3SR-TINY human evaluation. Each question is shown with the same ordered views given to the models; the interface supports option selection or numeric input, skipping, and revision, and never shows the ground truth. Tiny and full-set scores come from different question sets.

C.3 COMPLETE EVALUATION RESULTS

C.3.1 FULL MODEL EVALUATION RESULTS.

Tab. 10 reports all 32 evaluated models: 22 general open-source, seven spatially adapted, and three proprietary models. All values are percentages rounded to one decimal place. The results include five task families and all three scene splits. Incomplete runs are excluded.

Table 9: API model identifiers and evaluation request dates (UTC).

Model	ID	Request date
GPT-5.4	`gpt-5.4-2026-03-05`	Requested September 18, 2026.
Claude-Sonnet-5	`claude-sonnet-5`	Requested September 19, 2026.
Gemini-3.5-Flash	`gemini-3.5-flash`	Requested September 19, 2026.

C.3.2 Scaling with Model Size.

Fig. 24 compares full-set accuracy across model sizes for Object Counting, Relative Distance, and Relative Position XY/Z. Object Counting, Relative Distance, and Relative Position Z generally improve with scale, whereas Relative Position XY shows non-monotonic trends in some model families, as discussed in Sec. 4.2.

C.3.3 Multi-view Capacity.

Fig. 25 reports the current view-count intervention described in Sec. 5.3. Six open-source models are evaluated with one to six views. Object Counting and the three relative-relation tasks use 1,200 sampled QA pairs (400 per scene split). The shaded band represents one standard deviation across models. Results differ by task and scene split: in CONTEXTUAL, Relative Position Z improves with more views, whereas Absolute Distance largely saturates after three views, Relative Position XY shows little net gain, and Relative Distance declines from the first additional view.

D Diagnostic Analysis of Spatial Reasoning

D.1 Motivation: Full-View versus Single-View Evaluation

Fig. 1 in Sec. 1 compares full-view and single-view accuracy to measure how much a benchmark benefits from multi-view inputs. Let $\mathcal{Q}$ be the evaluated question set. Each question $q \in \mathcal{Q}$ has a view set $V_q = \{v_1, \ldots, v_{n_q}\}$ and ground-truth answer a_q, and $f(q, S)$ denotes the model prediction given the question and an ordered image set S. We define

$$\mathrm{Acc}_{\mathrm{Full}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}[f(q, V_q) = a_q], \qquad \mathrm{Acc}_{\mathrm{Single}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{n_q} \sum_{v \in V_q} \mathbb{1}[f(q, \{v\}) = a_q],$$
$$\mathrm{Acc}_{\mathrm{Blind}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}[f(q, \emptyset) = a_q], \qquad \Delta = \mathrm{Acc}_{\mathrm{Full}} - \mathrm{Acc}_{\mathrm{Single}}. \tag{4}$$

Single-view accuracy is thus an item-macro average: every view of a question is evaluated separately, and each question receives equal weight regardless of its view count. The Blind condition removes all images and measures answerability from text priors. A large positive Δ suggests that answering benefits from complementary evidence from multiple views, whereas $\Delta \approx 0$ or $\Delta < 0$ suggests limited empirical benefit from multi-view inputs, which may reflect either single-view sufficiency or a failure to integrate views. All conditions use the same prompt and greedy decoding; unparsed responses count as incorrect. Random is the expected accuracy of uniform guessing over each question's valid options. Only multiple-choice questions are used, so that all benchmarks share exact-match scoring: for View2Space and DM3SR (†), we use their multiple-choice subsets (View2Space: plain four-choice items).

Tab. 11 reports the complete results for Qwen3-VL-32B (Qwen Team, 2025) and GLM-4.5V (V Team et al., 2025), whose Δ values are plotted in Fig. 1. Full-view gains are small on MMSI-Bench ($+1.0$ and $+0.1$ pp) and View2Space ($+2.8$ and $+0.8$ pp), and full-view inputs reduce accuracy on MindCube for both models (-5.2 and -3.4 pp). In contrast, DM3SR yields the largest gains ($+8.4$ and $+9.5$ pp), and Full also exceeds Blind by 11.5 and 15.0 pp, indicating that its questions benefit substantially from multi-view evidence beyond what a single view or text priors provide.

Table 10: Complete DM3SR results for all 32 models evaluated on 4,775 questions (%). Absolute Distance uses MRA; all other tasks use exact-match accuracy. Green cells mark the top two overall scores within each model group, and gray cells mark the top two scores for each task and split within that group.

Model	Overall	By Task					By Scene Split		
	Avg.	Obj. Cnt.	Abs. Dist.	Rel. Dist.	Rel. Pos. XY	Rel. Pos. Z	CONTE-XTUAL	CONTROLLED SYN	CONTROLLED REAL
Proprietary Models (API)									
GPT-5.4	56.8	31.0	41.6	61.5	46.4	77.9	52.0	59.0	63.1
Gemini-3.5-Flash	61.7	41.5	52.1	64.4	46.3	85.6	56.4	64.9	66.2
Claude-Sonnet-5	55.7	20.9	46.4	58.1	51.0	74.7	52.9	54.8	67.1
Open-source Models									
Qwen2.5-VL-3B	43.3	10.0	39.9	51.8	41.8	47.4	44.4	40.9	48.5
Qwen2.5-VL-7B	47.6	20.1	38.6	54.2	44.1	56.9	44.9	48.3	53.1
Qwen2.5-VL-32B	48.6	19.4	44.5	59.2	40.8	54.8	45.4	48.4	58.3
Qwen2.5-VL-72B	48.6	17.1	39.7	55.4	40.3	63.8	43.7	50.0	57.8
Qwen3-VL-2B	36.2	4.7	34.6	43.5	37.1	37.9	38.6	31.9	43.9
Qwen3-VL-4B	45.2	18.5	39.0	54.2	42.5	49.0	40.5	46.9	53.2
Qwen3-VL-8B	46.5	24.4	26.5	59.5	39.8	54.1	40.7	51.0	47.9
Qwen3-VL-32B	55.8	34.4	37.5	64.5	50.3	66.7	49.3	59.6	61.4
InternVL3.5-2B	41.5	4.5	31.0	52.5	41.0	46.5	41.4	43.1	36.6
InternVL3.5-4B	44.3	10.2	39.0	53.9	40.9	50.4	43.5	44.6	46.1
InternVL3.5-8B	46.9	22.5	43.4	53.9	38.7	56.2	44.4	47.1	53.1
InternVL3.5-14B	45.8	22.7	31.9	55.1	39.8	55.0	41.0	47.5	53.9
InternVL3.5-38B	49.3	22.5	37.7	58.5	43.1	58.6	45.8	50.6	54.7
Ovis2.5-2B	38.5	2.4	38.4	43.7	42.0	41.6	37.9	38.2	41.2
Ovis2.5-9B	47.4	6.6	44.1	56.7	41.9	57.3	48.1	43.7	58.3
MiniCPM-V-4.5	42.8	13.3	36.7	45.4	42.7	53.4	40.5	43.3	47.7
GLM-4.5V	54.3	19.4	48.1	63.1	46.4	66.4	50.8	54.6	63.5
GLM-4.6V-Flash	47.5	15.6	46.7	56.3	47.5	48.1	47.7	44.7	56.8
Gemma-4-E2B	41.7	12.6	30.3	50.9	42.0	45.5	42.2	41.1	41.9
Gemma-4-E4B	44.1	12.3	26.3	52.4	40.4	57.5	43.2	44.4	45.7
Gemma-4-12B	48.2	14.7	37.1	53.8	42.5	64.2	46.0	49.2	51.0
Gemma-4-31B	57.2	38.6	46.8	60.5	45.6	76.3	51.6	59.4	65.9
Spatial Models									
SpaceR	47.8	18.5	36.2	54.1	46.2	57.7	46.6	47.5	52.3
ViLaSR	48.2	19.9	37.9	54.3	47.3	56.7	47.6	47.3	53.1
VG-LLM-8B	45.6	8.3	29.7	56.5	44.7	53.8	47.8	45.0	41.2
MindCube-RawQA	42.3	0.2	40.0	50.6	47.6	42.7	43.6	39.9	46.6
MindCube-CogMap	38.5	3.3	19.3	49.2	45.6	39.8	37.3	39.1	39.6
SpaceQwen2.5-VL-3B	34.2	6.2	23.7	47.9	22.6	43.0	31.8	34.5	40.1
SpaceThinker-Qwen2.5-VL-3B	42.1	10.4	28.2	51.4	42.0	48.6	41.2	43.0	41.6

Table 11: Full-view, single-view, and blind accuracy (%) across multi-view benchmarks (Eq. 4). Δ = Full − Single (pp); the largest Δ per model is in bold. †Multiple-choice subset.

Benchmark	N	Random	Qwen3-VL-32B				GLM-4.5V			
			Full	Single	Blind	Δ	Full	Single	Blind	Δ
MMSI-Bench (Yang et al., 2026)	1,000	25.0	33.0	32.0	32.0	+1.0	32.2	32.1	29.0	+0.1
MindCube (Yin et al., 2026)	1,050	33.3	33.0	38.2	42.5	-5.2	42.5	45.9	44.2	-3.4
View2Space† (Ke et al., 2026)	1,200	25.0	28.2	25.4	25.3	+2.8	27.9	27.1	26.8	+0.8
DM3SR†	3,775	44.5	61.2	52.8	49.7	**+8.4**	59.3	49.8	44.3	**+9.5**

D.2 CONTROL GROUP GENERATION DETAILS

D.2.1 MULTI-VIEW CAPACITY.

The experiment in Sec. 5.3 samples 400 QA pairs from each of the three scene splits, giving 1,200 pairs. Starting from one view, the input is varied through six views by randomly adding or removing one view at a time. Six open-source models (Qwen2.5-VL-32/72B, Qwen3-VL-32B, InternVL3.5-38B, GLM-4.5V, and Gemma-4-31B) are evaluated on the same sampled questions. Fig. 25 shows the mean over models, with one standard deviation as the shaded band. The paired question design isolates view-count changes from changes in question selection.

D.2.2 VIEWPOINT CONTINUITY.

The continuity intervention preserves each question's image set and answer while changing the order of images. For normalized camera-direction vectors $\mathbf{v}_i$, let $A_{ij} = \arccos(\mathrm{clip}(\mathbf{v}_i^\top \mathbf{v}_j, -1, 1))$. We choose the permutation with the shortest open angular path, thereby minimizing the mean angular deviation (MAD) defined in Eq. 2. Since a question contains at most six views, all permutations can be evaluated exactly. Ties are resolved by the fewest inversions relative to the original order and then lexicographically. Ordinal image references in questions and options are remapped to the new image positions. Tab. 12 summarizes the procedure.

Table 12: Exact viewpoint reordering used in the continuity control.

Step	Operation
1	Compute pairwise angular separations A_{ij} from the supplied camera directions.
2	Enumerate all permutations of the $N \leq 6$ supplied views and sum the $N-1$ consecutive angles.
3	Select the shortest path, breaking ties by inversion count and then lexicographic order.
4	Reorder images and remap all ordinal image references without changing the answer.

Over the 4,197 non-Absolute-Distance questions, reordering shortens the total consecutive angular path by 30.0% (30.3% on CONTROLLED-SYN, 30.9% on CONTEXTUAL, and 26.0% on CONTROLLED-REAL). One CONTROLLED-REAL Absolute Distance question references an unavailable camera pose and is excluded, so the paired evaluation contains 4,774 questions per model. Results are reported for eleven open-source models: Qwen2.5-VL-7/32/72B, Qwen3-VL-8B, InternVL3.5-14/38B, Ovis2.5-9B, Gemma-4-12/31B, SpaceR, and ViLaSR.

D.2.3 SPATIAL LAYOUTS.

The layout control in Sec. 5.1 uses 50 paired CONTROLLED-SYN scenes. Reducing normalized vertical spread from 1.0 to 0.5 changes the object-height standard deviation from 2.21 m to 1.18 m. Object identities, horizontal positions, camera views, question wording, and options remain fixed; answers are recomputed from the edited geometry. Twelve models answer the same 600 questions under each condition: 100 Object Counting, 200 Relative Distance, 150 Relative Position XY, and 150 Relative Position Z questions.

D.3 FULL EXPERIMENT RESULTS

Fig. 26 reports the task-level continuity comparison for the eleven models in Sec. 5.2, pooled over all scene splits. The mean overall change is $+1.40$ percentage points (pp): $+0.62$ pp on CONTEXTUAL and $+1.91$ pp on CONTROLLED (CONTROLLED-SYN and CONTROLLED-REAL combined). Relative Distance shows the largest mean gain ($+2.88$ pp; $+2.55$ and $+3.09$ pp on CONTEXTUAL and CONTROLLED), followed by Relative Position Z ($+1.46$ pp) and Relative Position XY ($+0.64$ pp), while Object Counting (-0.13 pp) and Absolute Distance (-0.06 pp) remain nearly unchanged. Other effects vary by model and task.

Tab. 13 reports the paired vertical-spread comparison. Compressing the vertical spread changes the twelve-model mean by $+1.71$ pp ($+3.1\%$ relative) for Relative Distance and $+1.94$ pp ($+3.5\%$) for Relative Position XY. Relative Position Z falls by -2.56 pp (-4.9%); Object Counting falls by -1.33 pp (-7.1%). Relative changes are computed as $(\text{compressed} - \text{original})/\text{original}$ using unrounded group means.

Table 13: Mean accuracy (%) over 12 models in the paired 50-scene layout experiment. The last two columns compare vertical-spread compression ($z = 0.5$) against the original ($z = 1.0$) using percentage-point and relative changes.

Task	$z = 1.0$	$z = 0.5$	Δ (pp)	Δ (relative)
Overall	48.50	48.69	+0.19	+0.4%
Object Counting	18.75	17.42	-1.33	-7.1%
Relative Distance	55.50	57.21	+1.71	+3.1%
Relative Position XY	54.89	56.83	+1.94	+3.5%
Relative Position Z	52.61	50.06	-2.56	-4.9%

E LICENSES AND LIMITATIONS

E.1 LICENSES

Tab. 14 lists the external assets used to build DM3SR. HSSD scenes are used for rendering with object class labels replaced (Appendix A); OmniObject3D scans and Poly Haven HDRIs are used without modification. Because HSSD permits only non-commercial use, DM3SR is released under CC BY-NC 4.0 for non-commercial research. Evaluated open-source models are used under their respective licenses, and proprietary models through their providers' APIs under the corresponding terms of service.

Table 14: External assets used in DM3SR.

Asset	License	Split	Use
HSSD (Khanna* et al., 2023)	CC BY-NC 4.0	CONTEXTUAL	Indoor scenes and object meshes
OmniObject3D (Wu et al., 2023)	CC BY 4.0	CONTROLLED-SYN	Scanned objects
Poly Haven HDRIs	CC0	CONTROLLED-SYN	Outdoor backgrounds and lighting

E.2 LIMITATIONS

Scene realism and scale. CONTROLLED-SYN and CONTEXTUAL are rendered, so results may not fully transfer to real imagery; CONTROLLED-REAL narrows this gap but contains only 48 scenes captured in a single room. **Generated labels.** CONTEXTUAL class labels are produced automatically from rendered views and were inspected on a sample, so residual labeling errors may remain. **Scope of diagnostics.** The controlled experiments cover subsets of models (6–12), and the vertical-spread intervention is applied only to 50 CONTROLLED-SYN scenes. **Task coverage.** DM3SR evaluates static geometric relations among objects and cameras; dynamic scenes, egocentric motion, and action-conditioned reasoning are not covered.

Contextual

(a) Absolute Distance **Answer: 4.03 m**



brown woven square basket — light wood six-drawer cabinet

Q: Estimate the 3D straight-line distance in meters between the centers of the brown woven square basket and the light wood six-drawer cabinet. The answer must be between 0.0 and 10.0 m.

(b) Relative Distance **Answer: B. white geometric patterned carpet**



wooden five-rung ladder — black flat screen television — white geometric patterned carpet

Q: Which object is closer to the wooden five-rung ladder?

A. black flat screen television B. white geometric patterned carpet

(c) Relative Position XY **Answer: D. front left**



white four-tier wire shelf cart — white all-in-one desktop computer — light gray patterned carpet

Q: Suppose you are standing at the location of the white four-tier wire shelf cart and facing white all-in-one desktop computer. Where is the light gray patterned carpet located relative to you?

A. front right B. back right C. back left D. front left

(d) Relative Position Z **Answer: C. brown metal ferris wheel model**



green potted bamboo — gray rectangular storage crate — brown metal ferris wheel model

Q: Which object's center is the highest?

A. green potted bamboo B. gray rectangular storage crate C. brown metal ferris wheel model

Figure 16: CONTEXTUAL QA examples. Colored wireframes mark the projected 3D bounding boxes of target objects (regardless of occlusion) for illustration only and are not part of the model inputs.

Controlled-Syn

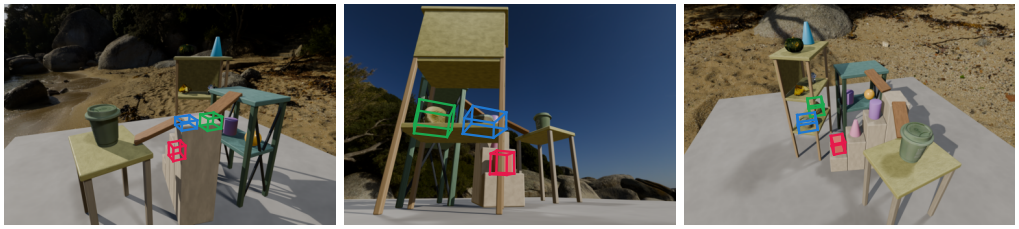
(a) Object Counting **Answer: 5**



pink cylinder orange box green sphere olive icosphere purple cone

Q: How many plain untextured objects are placed on the support structures?

(b) Relative Distance **Answer: B. ricecooker**

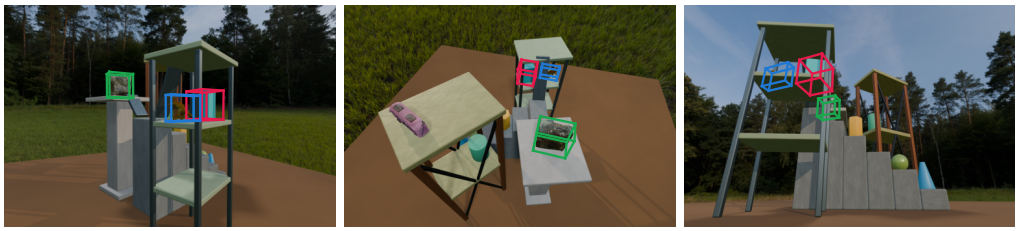


red box bun ricecooker

Q: Which object is closer to the red box?

A. bun B. ricecooker

(c) Relative Position XY **Answer: A. left**



cyan box mooncake dinosaur

Q: Suppose you are standing at the cyan box and facing the mooncake. Where is the dinosaur?

A. left B. right

(d) Relative Position Z **Answer: A. green cylinder**



pink box green cylinder candle

Q: Which object's center is the lowest?

A. green cylinder B. candle C. pink box

Figure 17: CONTROLLED-SYN QA examples. Colored wireframes mark the projected 3D bounding boxes of target objects (regardless of occlusion) for illustration only and are not part of the model inputs.

Controlled-Real

(a) Object Counting **Answer: 8**



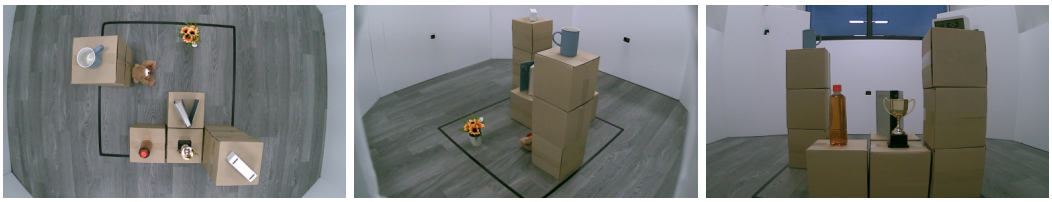
Q: How many objects are in the scene, excluding the cardboard box and the background?

(b) Absolute Distance **Answer: 34.64 cm**



Q: The cardboard box has an edge length of 20 cm. Estimate the 3D straight-line distance in centimeters between the centers of the toy car and the mug. The answer must be between 0 and 100 cm.

(c) Relative Distance **Answer: A. trophy**



Q: Which object is closer to the water bottle?

A. trophy B. flower pot

(d) Relative Position XY **Answer: A. left**



Q: Suppose you are standing at the location of the fan and facing toy car. Where is the headset located relative to you?

A. left B. right

Figure 18: CONTROLLED-REAL QA examples.

Numerical Task

You're helpful spatial-reasoning AI assistant. Given multi-view RGB images of a same 3D scene, you should solve the following task and answer the question based on the provided images.

{ *question* }

The final answer MUST BE single number, and should be put in `<answer></answer>` tags. Do not output anything other than the answer itself. In general, your response's format should be like `<answer>...</answer>`.

Figure 19: Evaluation prompt for numerical task.

Multi-Choice Task

You're helpful spatial-reasoning AI assistant. Given multi-view RGB images of a same 3D scene, you should solve the following task and answer the question based on the provided images.
{ *question* } { *options* }
The final answer MUST BE option's letter from choices, and should be put in `<answer></answer>` tags. Do not output anything other than the answer itself. In general, your response's format should be like `<answer>...</answer>`.

Figure 20: Evaluation prompt for multi-choice task.

Numerical Task (MindCube-RawQA)

[Task]
Your task is to analyze the spatial arrangement of objects in the scene by examining the provided images, which show the scene from different viewpoints.
[Answer Instruction]
You only need to provide *ONE* correct answer. Your final answer MUST BE single number. Your response should **only** be '`<answer>(replace with your answer)</answer>`'
[Question]
{ *question* }

Figure 21: Evaluation prompt for numerical task (MindCube-RawQA).

Multi-Choice Task (MindCube-RawQA)

[Task]
Your task is to analyze the spatial arrangement of objects in the scene by examining the provided images, which show the scene from different viewpoints.
[Answer Instruction]
You only need to provide *ONE* correct answer selecting from the options listed below. For example, if you think the correct answer is 'A. Above' from 'A. Above B. Under C. Front D. Behind', your response should **only** be '`<answer>A. Above</answer>`'
[Question]
{ *question* } { *options* }

Figure 22: Evaluation prompt for multi-choice task (MindCube-RawQA).

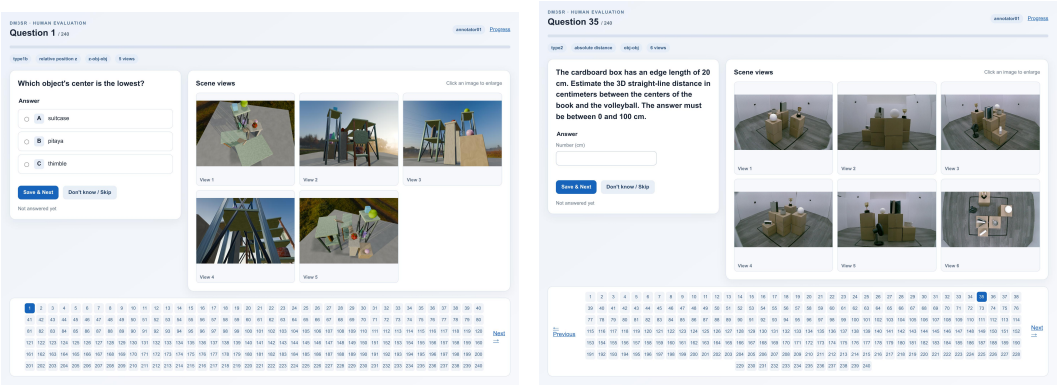


Figure 23: Web interface used for the DM3SR-TINY human evaluation.

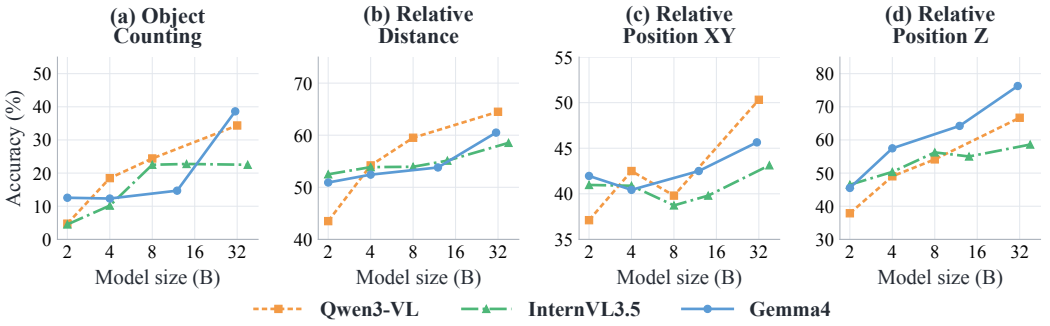


Figure 24: Full-set accuracy versus model size for Qwen3-VL, InternVL3.5, and Gemma4 on Object Counting, Relative Distance, Relative Position XY, and Relative Position Z (left to right). All four tasks use exact-match accuracy. Model size uses a logarithmic axis; each panel has its own accuracy range.

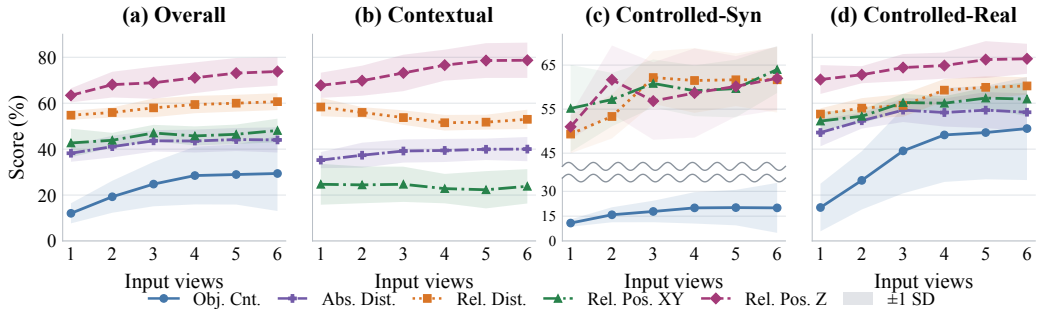
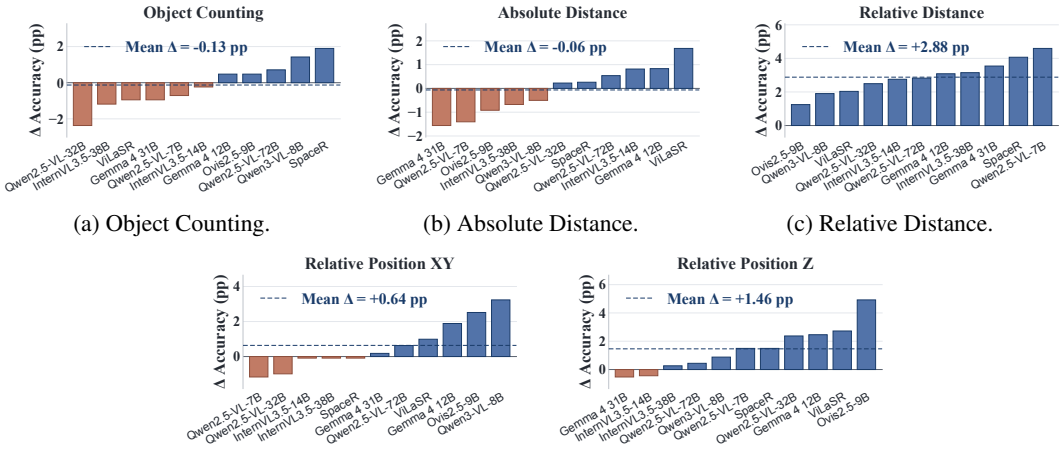


Figure 25: View-count ablation across six open-source models, shown for Overall, CONTEXTUAL, CONTROLLED-SYN, and CONTROLLED-REAL. Shading denotes one standard deviation across models. In (c), the y-axis is broken to show Object Counting (bottom) separately from the other tasks (top).



(a) Object Counting. (b) Absolute Distance. (c) Relative Distance.

(d) Relative Position XY. (e) Relative Position Z.

Figure 26: Task-level results for the viewpoint-continuity intervention. Differences are reordered minus original accuracy in percentage points; Absolute Distance uses MRA.