DISENTANGLING 3D MODELING FROM SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we explore an alternative paradigm for spatial reasoning by explicitly disentangling 3D perception from reasoning, rather than jointly acquiring implicit 3D perception and reasoning through large-scale training. Our key observation is that modern perception models excel at estimating continuous 3D geometry, whereas large language models (LLMs) are particularly effective at compositional and symbolic reasoning. Motivated by these complementary strengths, we propose the *Disentangled Spatial Reasoner (DiSR)*, a simple yet effective framework that reconstructs the physical world into structured 3D evidence using off-the-shelf expert perception models and fine-tunes an LLM with LoRA to perform reasoning solely over this explicit geometric evidence. Without large-scale 3D VQA training or complex tool-use policies, *DiSR* achieves competitive performance on popular spatial reasoning benchmarks. Beyond its strong performance, *DiSR* offers improved interpretability, modularity, and computational efficiency, demonstrating that explicit separation of perception and reasoning is a scalable and effective alternative paradigm to end-to-end modeling for spatial intelligence.

1 INTRODUCTION

Spatial reasoning is a fundamental capability for intelligent systems to understand, interpret, and interact with the physical 3D world. Unlike conventional visual recognition, which focuses on identifying objects and their semantic attributes, spatial reasoning requires models to infer geometric properties, spatial relationships, and physical constraints from visual observations. Such reasoning is essential in a wide range of applications, including embodied agents, robotic manipulation, autonomous driving, augmented reality, and human-robot interaction, where decision-making depends not only on *what* objects are present but also on *where* they are and *how* they relate to one another in 3D space. As multimodal large language models (MLLMs) continue to demonstrate strong progress in visual understanding and language reasoning (Liu et al., 2023; OpenAI, 2023; Team, 2025), enabling reliable spatial reasoning has become a key step toward building AGI systems.

Despite these advances, achieving robust spatial intelligence with MLLMs remains challenging. Existing approaches either train MLLMs on large-scale 3D visual question answering (3D VQA) datasets to acquire geometric understanding (Chen et al., 2024; Cheng et al., 2024; Ma et al., 2025c;b; Liang et al., 2026), or align explicit 3D latent representations with MLLMs through multimodal learning (Hong et al., 2023; Zhu et al., 2023b; Xu et al., 2024; Huang et al., 2024) as shown in Figure 1(a). These approaches share a common formulation: jointly optimizing implicit geometric perception and spatial reasoning. Consequently, 3D knowledge is embedded in latent representations rather than explicitly reconstructed into interpretable geometric structures. Moreover, such large-scale training often requires substantial computational resources, limiting the scalability and efficiency of spatial reasoning systems. Recently, agentic frameworks shown in Figure 1(b) have explored the use of external tools to mitigate these challenges, but introduce additional complexity by requiring MLLMs to learn effective tool-use strategies through reinforcement learning or strong foundation models (Ma et al., 2024; Chen et al., 2026a; Han et al., 2025; Chen et al., 2026b).

Disentangling 3D Perception from Spatial Reasoning. Rather than further coupling perception and reasoning, we explore an alternative paradigm based on explicit disentanglement. This idea is motivated by the complementary strengths of modern perception models and large language models (LLMs): specialized perception models have demonstrated strong capabilities in estimating con-

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.

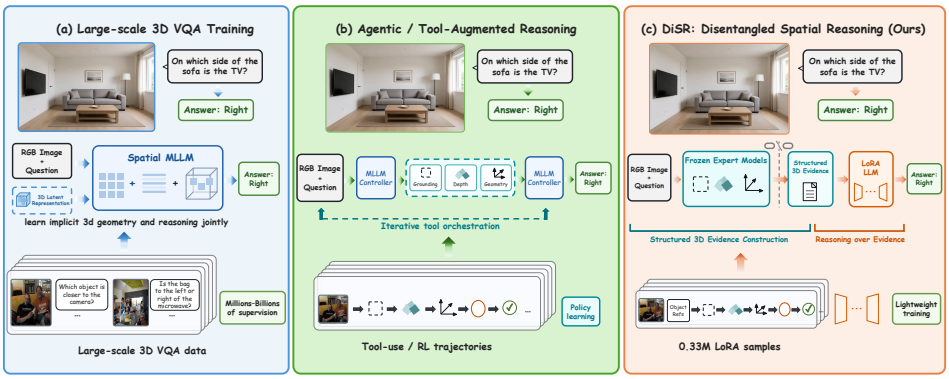


Figure 1: **Comparison of spatial reasoning paradigms.** (a) Methods jointly acquire implicit geometric perception and spatial reasoning through extensive spatial supervision, whereas (b) tool-augmented agents learn tool-use policy and invoke external tools through iterative execution. (c) *DiSR* instead constructs explicit 3D evidence and then performs reasoning with an LLM solely on the structured evidence, thereby disentangling 3D modeling from spatial reasoning.

tinuous geometric properties, such as depth, object poses, and 3D layouts, while LLMs excel at compositional and symbolic reasoning over structured information. These observations raise a natural question: can spatial intelligence be achieved by allowing each component to specialize in its respective capability, rather than requiring a single MLLM to jointly learn both?

Based on this insight, we propose to explicitly separate 3D perception from spatial reasoning as shown in Figure 1(c). Specifically, we leverage off-the-shelf expert perception models to reconstruct the physical world into structured 3D evidence, including 3D locations, object orientations and sizes. An LLM with LoRA is then fine-tuned to reason solely over this explicit geometric evidence. By introducing an interpretable intermediate textual evidence between perception and reasoning, our method offers a simple, efficient, and modular alternative to end-to-end joint modeling.

Disentangled Spatial Reasoner (DiSR). To instantiate this idea, we propose the *Disentangled Spatial Reasoner (DiSR)*, illustrated in Figure 1(c). *DiSR* decomposes spatial reasoning into two steps: (1) *structured 3D evidence construction*, where specialized perception models produce structured geometric evidence, including object 3D locations, object sizes, and object orientations; and (2) *reasoning over structured 3D evidence*, where an LLM performs structured reasoning solely over the extracted geometric evidence inputs without images to produce the final answer.

This disentangled design yields several practical advantages. First, by offloading geometric perception to specialized perception models, *DiSR* significantly reduces the need for large-scale 3D supervision during MLLM training, leading to more efficient training and lower computational cost, as summarized in Table 1. Second, the modular formulation improves interpretability and diagnosability: since structured 3D evidence is explicitly available, reasoning errors can be clearly separated from perception errors, as demonstrated by diagnostic analysis in Table 6. Third, the framework is highly extensible, allowing improvements in either perception models or LLMs to be incorporated independently without retraining the full system, as evidenced by results in Table 7.

Experimental Results. We evaluate *DiSR* on three representative benchmarks, including 3DSR-Bench (Ma et al., 2025a), SPAR-Bench (Zhang et al., 2025), and CV-Bench-3D (Tong et al., 2024). Despite using significantly less computational cost and fewer trainable parameters, *DiSR* consistently outperforms prior methods that rely on large-scale 3D training, such as SpatialRGPT (Cheng et al., 2024) and SpatialReasoner (Ma et al., 2025b). Specifically, we achieve new state-of-the-art performance on 3DSRBench and SPAR-Bench, outperforming previous best methods by **3.77%** and **1.70%**, respectively. These results demonstrate that explicitly disentangling 3D perception from reasoning is an effective and data-efficient paradigm for spatial intelligence.

Overall, our contributions are summarized as follows:

- We propose to disentangle 3D perception from spatial reasoning and introduce *DiSR*, which leverages 3D perception expert models to construct explicit evidence and reason solely over the structured evidence with LLMs.

- We demonstrate that *DiSR* with this disentangled design achieves strong performance while using significantly less training data and computational resources.
- We validate *DiSR* on popular spatial reasoning benchmarks, where it consistently outperforms prior large-scale training approaches, achieving new state-of-the-art results on the challenging 3DSRBench and SPAR-Bench.

2 Related Work

Spatial VLMs with Large-scale 3D Supervision. A dominant approach to spatial reasoning is to enhance MLLMs through large-scale 3D supervision. Early works construct 3D visual question answering (3D VQA) datasets from monocular images and fine-tune VLMs to implicitly acquire geometric knowledge. For example, SpatialVLM (Chen et al., 2024) introduces Internet-scale 3D VQA data, while SpatialRGPT (Cheng et al., 2024) adopts region-level 3D scene graphs and depth cues to improve geometric representation learning. Subsequent studies, including SpatialLLM (Ma et al., 2025c), SpatialReasoner (Ma et al., 2025b), and HiSpatial (Liang et al., 2026), explore improved 3D-aware data, training strategies, reinforcement learning, and RGB-D representations to further enhance spatial reasoning capabilities.

Despite their differences, these methods share a common formulation: jointly learning implicit 3D perception and spatial reasoning within a single MLLM through large-scale optimization. This design often requires substantial 3D supervision and computational resources, while coupling perception and reasoning into an implicit representation makes error diagnosis challenging.

Aligning MLLMs with 3D Representations. Another line of research equips MLLMs with spatial intelligence by aligning them with explicit 3D representations. Instead of relying solely on 3D VQA supervision, these methods bridge the modality gap between geometric representations and language models through dedicated encoders, alignment objectives, or instruction tuning. For example, 3D-LLM (Hong et al., 2023) projects 3D scene features into the language space to support tasks such as question answering, grounding, navigation, and dialogue. 3D-VisTA (Zhu et al., 2023b) pre-trains unified 3D vision-language representations over RGB-D scenes and scene-text pairs for a variety of downstream tasks. PointLLM (Xu et al., 2024) aligns point-cloud encoders with LLMs using point-text instruction tuning, enabling language models to understand object-level 3D geometry. LEO (Huang et al., 2024) further extends this paradigm to embodied agents through joint 3D vision-language alignment and vision-language-action instruction tuning. These methods aim to bridge the modality gap between 3D representations and MLLMs, which suffer from similar problems with methods training on large-scale 3D supervision.

Agentic Spatial Reasoning. Another line of research enhances spatial reasoning by augmenting MLLMs with external perception models, geometric tools, or formal reasoning modules. Instead of encoding all geometric knowledge within the language model, these approaches enable MLLMs to acquire spatial capabilities through tool invocation. For example, SpatialPIN (Ma et al., 2024) improves spatial reasoning in a training-free manner by leveraging priors from multiple 3D foundation models. SpaceTools (Chen et al., 2026a) trains MLLMs to coordinate multiple perception and robotics tools through reinforcement learning, while TIGeR (Han et al., 2025) equips MLLMs with geometric computation tools for precise spatial operations. GCA (Chen et al., 2026b) further incorporates formal constraints to connect semantic understanding with geometric reasoning. These methods demonstrate the effectiveness of external geometric computation for spatial reasoning. However, they require learning tool-use policies, orchestrating multi-step interactions, or designing task-specific constraints.

3 Method

Modern perception models have demonstrated remarkable capabilities in estimating continuous geometric properties, whereas LLMs excel at compositional and symbolic reasoning. Motivated by these complementary strengths, we propose the *Disentangled Spatial Reasoner (DiSR)*, which formulates spatial reasoning as a two-stage process: reconstructing the physical world into structured 3D evidence, followed by reasoning over this explicit geometric evidence. Rather than jointly optimizing perception and reasoning, *DiSR* explicitly separates the two through an interpretable inter-

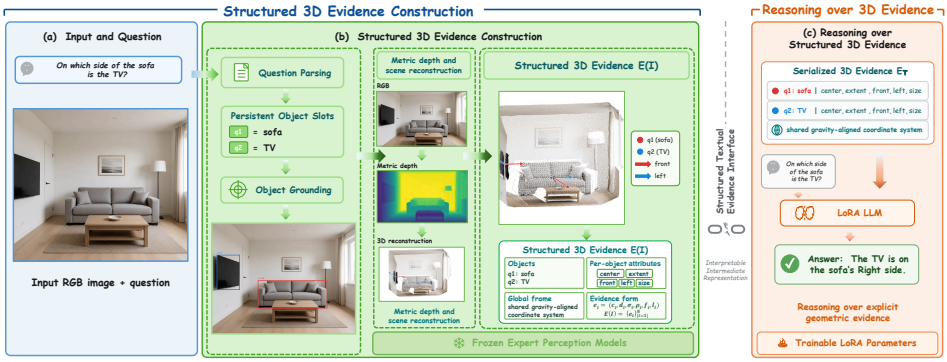


Figure 2: **Overview of DiSR.**

mediate textual evidence, resulting in a simple, modular, and efficient framework. We first present an overview of DiSR and then describe each component in detail.

3.1 Overview of DiSR

As illustrated in Figure 2, the key idea of *DiSR* is to explicitly separate 3D perception from spatial reasoning through structured intermediate 3D evidence. Instead of requiring a single MLLM to jointly learn geometric perception and reasoning from large-scale 3D VQA data, *DiSR* assigns each capability to a specialized component: expert perception models are responsible for constructing structured 3D evidence, while an LLM focuses on reasoning solely over the resulting geometric representation.

Given an input image I and a spatial reasoning question Q, *DiSR* predicts the answer A through two stages: (1) structured 3D evidence construction, and (2) reasoning over 3D evidence. Specifically, a question parser first identifies the relevant objects required for answering Q, and off-the-shelf perception models are employed to estimate object-level geometric information from I, including 3D locations, object orientations, and object sizes. Then, the extracted geometric evidence is converted into a structured textual representation and provided to an LLM, which performs spatial reasoning and generates the final answer.

Formally, the overall process can be formulated as:

$$A = f_\theta(E(I), Q), \tag{1}$$

where $E(I)$ denotes the structured 3D evidence reconstructed from the image, and f_θ represents the reasoning model parameterized by θ. Unlike end-to-end methods that jointly optimize perception and reasoning, *DiSR* explicitly exposes the structured geometric evidence, enabling each component to specialize in its corresponding capability.

3.2 Structured 3D Evidence Construction

To initiate *DiSR*, we reconstruct the physical scene into structured 3D evidence. Instead of learning implicit geometric representations within the MLLM, we leverage perception expert models specialized for 3D geometric estimation.

Question Parsing. The question parsing identifies the objects required for answering the question and establishes the interface between language understanding and 3D perception. Given a question Q, we use Qwen3-VL-8B-Instruct as an object planner $f_{\text{plan}}(\cdot)$ to generate an object-centric plan:

$$P = f_{\text{plan}}(Q) = \{(q_i : m_i)\}_{i=1}^{N}, \tag{2}$$

where each tuple contains a persistent object slot q_i and its corresponding object reference m_i. The slot identifier uniquely tracks the same object throughout the subsequent evidence extraction and reasoning, while the object reference specifies the object to be localized in the image. For example, a question involving a red chair and a table produces:

$$\{\texttt{q1} : \text{“the red chair”}, \quad \texttt{q2} : \text{“table”}\}. \tag{3}$$

Object Grounding. Given the object plan P, we take Qwen3-VL-8B-Instruct as the grounding module to localize each queried object reference m_i in the image and identify its corresponding region r_i. The grounded regions are represented as:

$$R = \{(q_i : r_i)\}_{i=1}^{N}, \tag{4}$$

where the slot identifier q_i preserves the correspondence between the language query and the visual region. It bridges linguistic references, such as "the chair on the left" or "the object highlighted by the red box", with localized visual targets. By retaining only the regions specified in P, ***DiSR*** performs target-specific perception and avoids unnecessary computation cost on uninterested objects.

Reconstructing the Physical Scene to Derive Evidence. Given the grounded regions R, ***DiSR*** uses frozen expert perception models to recover geometric information. Specifically, SAM (Kirillov et al., 2023) extracts object masks, Metric3D (Yin et al., 2023) estimates metric depth, and WildCamera (Zhu et al., 2023a) estimates intrinsic parameters and uses PerspectiveFields (Jin et al., 2023) to recover gravity-related camera orientation. The depth map is then back-projected and transformed into a shared, gravity-aligned coordinate frame. For each grounded mask region, ***DiSR*** extracts the associated 3D points, removes unreliable outliers, and computes an object-centric 3D extent and center. In addition, OrientAnything (Wang et al., 2025) estimates object orientations, which are transformed into the same coordinate frame.

The resulting object-centric scene representation is defined as:

$$E(I) = \{e_i\}_{i=1}^{N}, \quad e_i = (\mathbf{c}_i, \mathbf{d}_i, \sigma_i, \rho_i, \mathbf{f}_i, \mathbf{l}_i), \tag{5}$$

where $\mathbf{c}_i \in \mathbb{R}^3$ denotes the object center, $\mathbf{d}_i \in \mathbb{R}^3$ denotes its 3D extent, σ_i and ρ_i represent object-level size descriptors, $\mathbf{f}_i$ and $\mathbf{l}_i$ denote the estimated front and left axes. Since all objects are represented in a shared gravity-aligned coordinate system, spatial comparisons can be performed directly from these geometric primitives without requiring the reasoner to interpret raw visual signals, thus eliminating large-scale 3D VQA training to learn implicit 3D geometry.

3.3 Reasoning over Structured 3D Evidence

After obtaining structured 3D evidence E, ***DiSR*** employs an LLM with LoRA to perform high-level spatial reasoning. The objective is not to learn geometric perception from raw visual inputs, but to infer spatial relationships and answer questions based on the explicit geometric evidence.

To bridge the perception output and the language model, the structured evidence E is serialized into a textual description with language:

$$E_T = \phi(E(I)), \tag{6}$$

where $\phi(\cdot)$ denotes the evidence serialization function. The resulting sequence contains object identities and their geometric attributes, which can be directly interpreted by the LLM. Given the question Q and serialized evidence E_T, the LLM generates the answer:

$$A = f_\theta(E_T, Q). \tag{7}$$

Since the geometric information is explicitly provided, the LLM only needs to learn reasoning patterns over the structured 3D evidence. We adopt LoRA-based fine-tuning to efficiently adapt the pretrained LLM while preserving its general reasoning capability.

3.4 Data Collection and Training

Training Data Construction. Following previous work (Ma et al., 2025b), we construct the training dataset from images in Open Images (Kuznetsova et al., 2018) by generating spatial reasoning questions and corresponding answers, resulting in a dataset $\mathcal{D} = \{(I, Q, A)\}$. Since ***DiSR*** performs question parsing for evidence extraction and spatial reasoning as two sequential stages during inference, we decompose each training sample into two complementary supervision tasks. Specifically, the question parsing task learns to predict the parsing result (Q, P). The reasoning task learns to generate the answer conditioned on the extracted 3D evidence, (Q, E_T, A). This decomposition enables the LLM to jointly learn accurate question parsing and reasoning over explicit geometric evidence. Since images are not required during training, we also construct synthetic data that approximates the distribution of the collected dataset.

Table 1: **Results on 3DSRBench and efficiency comparisons.** Training efficiency is measured by GPU hours. "NR" indicates that the number is not reported. "*" represents additional data regarding "Relational Selection" task is used.

Method	Overall	Height	Location	Orientation	Multi-Object	Spatial Data	Reported Training
Open-source generalist models							
LLaVA-NeXT-8B	48.40	50.60	59.90	36.10	43.40	–	–
Qwen2.5-VL-7B-Instruct	48.40	44.10	62.70	40.60	40.50	–	–
Qwen3-VL-8B-Instruct	53.70	48.00	71.00	40.50	46.70	–	–
Qwen2.5-VL-72B-Instruct	54.90	53.30	71.00	43.10	46.60	–	–
Proprietary models							
GPT-4o	44.20	53.20	59.60	21.60	39.00	NR	NR
Claude-3.5-Sonnet	48.20	53.50	63.10	31.40	41.30	NR	NR
Gemini-2.0-Flash Thinking	51.10	53.00	67.10	35.80	43.60	NR	NR
QwenVLMax	52.00	45.10	70.70	37.70	44.80	NR	NR
Spatial specialist models							
SpaceLLaVA	42.00	49.30	54.40	27.60	35.40	NR	NR
SpatialBot	41.00	40.40	54.40	31.90	33.50	NR	A100 / 120h
SpatialLLM	44.80	45.80	61.60	30.00	36.70	NR	A100 / NR
SpatialRGPT w/ depth	48.40	55.90	60.00	34.20	42.30	NR	A100 / 256h
SpatialReasoner	60.30	52.50	**75.20**	55.20	51.80	0.36M	H100 / NR
HiSpatial-3B (RGB-XYZ)	63.85	**70.29**	66.17	**76.76**	51.20	15.68M	H100 / 1536h
DiSR-8B-LoRA	**67.62**	62.75	72.85	69.40	**63.30**	0.33M	**RTX 4080S / 59h**
DiSR-8B-LoRA*	65.81	62.61	69.81	67.86	61.90	0.36M	**RTX 4080S / 67h**

Parameter-Efficient Fine-tuning. We adopt LoRA to efficiently fine-tune the reasoning LLM while keeping all expert perception models frozen. Given the constructed training data, the LLM is optimized using the standard autoregressive language modeling objective:

$$\mathcal{L} = -\sum_{t=1}^{|Y|} \log p_\theta(y_t \mid y_{<t}, X), \tag{8}$$

where X denotes the corresponding input prompt (i.e., the parsing or reasoning task), and Y denotes the target output sequence. During training, only the LoRA parameters are updated, resulting in an efficient adaptation of the pretrained LLM without modifying the perception modules.

4 Experiments

To validate the effectiveness of our *DiSR*, we evaluate our model on: 1) three spatial reasoning benchmarks; 2) five general reasoning benchmarks. In this section, we first introduce our experimental setup 4.1, then present results on spatial reasoning 4.2 and general reasoning benchmarks 4.3, followed by ablation study 4.4 for disentangled error analysis of perception and reasoning.

4.1 Experimental Setup

Evaluation Benchmarks. We evaluate *DiSR* on three representative spatial reasoning benchmarks. 3DSRBench (Ma et al., 2025a) provides a fine-grained evaluation of 3D spatial reasoning across height, location, orientation, and multi-object relations. CV-Bench-3D (Tong et al., 2024) evaluates qualitative depth and distance reasoning beyond two-dimensional image-plane cues. SPAR-Bench (Zhang et al., 2025) covers tasks of depth estimation, metric distance estimation, and relational selection. The numerical tasks are evaluated using mean relative accuracy (MRA), while the relational-selection tasks are evaluated using accuracy.

Moreover, we examine whether the fine-tuned *DiSR* models retain strong general visual-reasoning capabilities on MMBench (Liu et al., 2024), GQA (Hudson & Manning, 2019), POPE (Li et al., 2023b), SEED (Li et al., 2023a), and RealWorldQA (x.ai, 2024). Finally, we evaluate object-grounding performance on CV-Bench-3D.

Comparison Methods. We compare *DiSR* with three groups of representative baselines: (1) general-purpose open-source MLLMs, including LLaVA, Cambrian, and the Qwen-VL family; (2) proprietary MLLMs, including GPT-4o, Claude, Gemini, and QwenVLMax; and (3) spatially specialized models, including SpatialLLM, SpatialRGPT, SpatialReasoner, and HiSpatial.

Table 2: **Results on SPAR-Bench.** "*" represents additional data regarding "Relational Selection" task is used.

Model	Metric Estimation (MRA)					Relational Selection (Acc.)			Avg.
	Depth-OC	Depth-OO	Dist-OC	Dist-OO	Metric Avg.	DistI-OO	ObjRel-OO	Rel. Avg.	
Qwen2.5-VL-7B-Instruct	32.14	20.27	42.75	25.37	30.13	58.82	52.47	55.65	38.64
Qwen3-VL-8B-Instruct	47.06	17.80	19.08	50.25	33.55	72.35	61.26	66.81	44.63
SpatialReasoner	29.86	17.77	21.40	40.55	27.40	15.29	16.21	15.75	23.51
HiSpatial-3B (RGB-XYZ)	41.41	14.15	7.68	**53.17**	29.10	**81.18**	59.07	**70.12**	42.78
DiSR-8B-LoRA	57.20	27.35	**67.66**	45.26	49.37	54.12	26.37	40.25	46.33
DiSR-8B-LoRA*	**64.82**	**27.37**	63.29	**53.27**	**52.19**	67.65	**68.96**	68.30	**57.56**

Table 3: **Results on CV-Bench-3D.** "w/ GT grounding" replaces model-predicted object boxes with ground-truth boxes to simulate user-specified targets.

Method	Depth ↑	Distance ↑	Avg. ↑
Qwen2.5-VL-7B-Instruct	84.50	80.00	82.25
Qwen3-VL-8B-Instruct	95.50	89.83	92.67
SpatialReasoner	87.00	72.83	79.92
HiSpatial-3B (RGB)	-	-	95.58
DiSR-8B-LoRA	92.83	91.67	92.25
DiSR-8B-LoRA w/ GT grounding	94.17	93.33	93.75

Implementation Details. We instantiate *DiSR* with Qwen3-VL-8B-Instruct, denoted as *DiSR-8B-LoRA*. Specifically, for object grounding in deriving 3D evidence, we use both the vision encoder and LLM backbone, while only the LLM is adopted for reasoning and question parsing. All expert perception modules remain frozen throughout training. We fine-tune the LLM by applying LoRA with rank 128, scaling factor 256, a learning rate of $2e-5$, and a maximum sequence length of 1024. A total of 0.33M training data is used. As *DiSR* doesn't need input images during training, a single RTX 4080 SUPER can provide enough memory to support 59 GPU hours of training.

4.2 Performance on Spatial Reasoning Benchmarks

Results on 3DSRBench. Table 1 compares 3DSRBench performance. *DiSR-8B-LoRA* achieves an overall accuracy of 67.62%, significantly outperforming the strongest prior method HiSpatial by **3.77%**. We observe that improvements are more pronounced on multi-object reasoning, where *DiSR* models obtain 63.30%, exceeding HiSpatial by about **12.1%** and SpatialReasoner by about **11.5%** individually. These results indicate that complex reasoning over explicit 3D evidence is particularly effective for relations that require combining geometric information across multiple objects.

Results on SPAR-Bench. Table 2 reports performance comparisons on SPAR-Bench. *DiSR-8B-LoRA* achieves the best overall score of 46.33%, outperforming Qwen3-VL-8B-Instruct by **1.70%** and the strongest prior spatial model HiSpatial by **3.55%**. An interesting phenomenon is observed: *our DiSR achieves 49.37% for metric estimation tasks, significantly surpassing Qwen3-VL-8B-Instruct and HiSpatial by **15.82%** and **20.27%**, respectively. However, DiSR performs poorly on relational selection tasks.* The reason behind this is that we follow SpatialReasoner for the training data construction, and the generated 0.33M training data can not cover the data distribution required for the relational selection tasks. With an additional 2000 data items regarding the "Relational Selection" task, the performance of *DiSR* is significantly enhanced, surpassing the previous best model by **12.93%** overall score.

Results on CV-Bench-3D. Table 3 reports the results on CV-Bench-3D. *DiSR* only achieves comparable performance to Qwen3-VL-8B-Instruct. With a deep analysis, we observe the two important reasons: (I) as identified by previous work Ma et al. (2025b), CV-Bench-3D contains substantial 2D shortcuts, which allow end-to-end MLLMs to infer answers without knowing the 3D geometry. (II) CV-Bench-3D suffers from low-quality images. It is challenging for current perception models to reconstruct accurate 3D evidence for small objects in low-quality images. Please refer to the Appendix for more detailed analysis.

The contrast between CV-Bench-3D accuracy in Table 3 and the grounding results in Table 5 shows a surprising phenomenon: *while HiSpatial achieves strong performance on CV-Bench-3D, it cannot*

Table 4: **Results on general multimodal reasoning benchmarks.**

Model	MMBench	GQA	POPE	SEED	RealWorldQA
HiSpatial-3B (RGB-XYZ)	69.67	61.50	87.97	63.51	58.95
SpatialReasoner	79.38	61.80	85.54	73.87	66.14
Qwen2.5-VL-7B-Instruct	84.02	61.23	87.61	77.64	68.76
Qwen3-VL-8B-Instruct	84.71	61.44	88.77	78.55	69.02
DiSR-8B-LoRA	84.62	61.20	88.72	78.95	68.76

even infer reliable object localization, leaving a mystery about how spatial reasoning performs in the model. In contrast, *DiSR* explicitly reconstructs structured 3D evidence and performs reasoning over this intermediate representation, making the geometric reasoning process inspectable and allowing errors to be attributed to grounding, 3D modeling, or reasoning.

In interactive applications, users can directly specify the queried objects instead of relying on model-predicted grounding. To simulate this setting, we replace the predicted object boxes with ground-truth boxes while keeping the same DiSR-8B-LoRA checkpoint and reasoning pipeline. As shown in Table 3, the average accuracy increases from 92.25% to 93.75%, with improvements of 1.34% and 1.66% on the Depth and Distance tasks, respectively. This controlled improvement identifies object grounding as a remaining performance bottleneck and demonstrates that stronger grounding models or user-provided object specifications can be incorporated for enhancement.

Training Efficiency Discussion. *DiSR* is highly training-efficient. Specifically, it only fine-tunes the LoRA parameters using 0.33M training samples on a single RTX 4080 SUPER GPU for 59 hours. In contrast, HiSpatial fine-tunes the entire MLLM on 2 billion spatial QA pairs using 32×48 H100 GPU hours. Despite this substantial reduction in training data, trainable parameters, and computational cost, *DiSR* achieves superior spatial reasoning performance. These results suggest that strong spatial reasoning does not necessarily require jointly learning 3D perception and reasoning through large-scale spatial supervision. Instead, by explicitly disentangling the two, specialized perception models can reconstruct reliable geometric evidence, allowing the LLM to focus on high-level compositional reasoning over structured 3D representations.

4.3 Performance on General Reasoning Benchmarks

Spatial reasoning fine-tuning might improve domain-specific performance at the expense of the general visual understanding capabilities inherited from the pretrained MLLM. To evaluate this trade-off, we assess *DiSR* on five general visual reasoning benchmarks. As shown in Table 4 , *DiSR* consistently achieves performance comparable to its base model, *i.e.*, Qwen3-VL-8B-Instruct, indicating that the substantial gains on spatial reasoning are achieved while preserving the general reasoning capabilities of the pretrained MLLMs.

In contrast, prior methods exhibit less favorable trade-offs. SpatialReasoner suffers a noticeable drop in general visual reasoning after fine-tuning on a relatively small 3D VQA dataset. HiSpatial is optimized on large-scale 3D VQA data and even improves over its base MLLM on general visual reasoning benchmarks. However, as shown in Table 5, both HiSpatial and SpatialReasoner experience a substantial decline in object grounding accuracy. These results demonstrate that the explicit disentanglement of perception and reasoning enables *DiSR* to specialize in spatial reasoning while retaining the broad visual reasoning capabilities inherited from the pretrained MLLM.

4.4 Ablation Study

Interpretability and Diagnosability of *DiSR*. By explicitly disentangling perception from reasoning, *DiSR* provides an interpretable reasoning pipeline in which errors can be attributed to individual components rather than an opaque end-to-end model. Since CV-Bench-3D provides ground-truth annotations for the intermediate 3D information required by *DiSR*, we conduct a diagnosability analysis to identify the performance bottlenecks.

Table 5: **Object grounding on COCO.**

Model	mIoU ↑	IoU@0.50 ↑
Qwen2.5-VL-7B-Instruct	64.02	71.64
Qwen3-VL-8B-Instruct	71.05	78.81
SpatialReasoner	34.82	40.51
HiSpatial-3B (RGB-XYZ)	fails	–
DiSR-8B-LoRA	**73.99**	**81.84**

Table 6: **Interpretability and diagnosability of DiSR on CV-Bench-3D.**

Method	Depth	Distance
DiSR-8B-LoRA	92.83	91.67
w/ gt of question parsing	92.33	92.83
w/ gt of object grounding	94.17	93.33
w/ gt of required 3D evidence	**100.0**	**100.0**

Table 7: **Extensibility of DiSR.**

Base MLLM	Height	Location	Orientation	Multi-Object	Overall
Qwen2.5-VL-7B-Instruct	60.43	71.80	64.96	61.67	65.52
Qwen3-VL-8B-Instruct	62.75	72.85	69.40	63.30	67.62

Table 6 presents the results. Replacing the predicted question parsing with ground truth yields only marginal improvements, indicating that question parsing is not the primary source of error. In contrast, using ground-truth object grounding improves the accuracy of the Depth and Distance tasks by 1.34% and 1.66%, respectively, suggesting that more accurate object grounding—potentially enabled by stronger MLLMs—would further improve overall spatial reasoning performance. Finally, when we replace the required 3D information with ground truth, *DiSR* achieves nearly 100% accuracy on both Depth and Distance, showing that the reasoning module can reliably infer spatial relationships when given accurate 3D evidence.

Object Grounding Analysis. Table 5 reports category-conditioned object grounding on 8,548 unambiguous examples constructed from COCO val2017 Lin et al. (2014), retaining only image–category pairs with exactly one valid instance. Although *DiSR* fine-tunes the LLM with LoRA for question parsing and reasoning, DiSR-8B-LoRA preserves and slightly improves the grounding capability of Qwen3-VL-8B, achieving higher valid-box coverage and localization accuracy under the same semantic grounding instruction. In contrast, SpatialReasoner exhibits substantially lower IoU despite returning valid boxes for most examples while HiSpatial even fails to return valid boxes. This discrepancy suggests that strong downstream spatial reasoning does not necessarily imply reliable object grounding, while the lightweight adaptation used by *DiSR* retains the visual localization interface required for structured evidence construction.

Cross-backbone Evaluation. We evaluate the extensibility of *DiSR* by instantiating it with different backbone MLLMs. Besides Qwen3-VL-8B-Instruct, we adopt Qwen2.5-VL-7B-Instruct as the base MLLM. As shown in Table 7, *DiSR* consistently achieves strong spatial reasoning performance across both backbones, demonstrating that the framework can seamlessly incorporate advances in foundation models without modifying its overall design. Moreover, a stronger backbone further improves spatial reasoning performance, primarily due to its enhanced general reasoning ability (Table 4) and more accurate object grounding capability (Table 5). These results highlight the modularity of *DiSR*, where improvements in the underlying MLLM can be directly translated into stronger spatial reasoning performance.

5 Conclusion

We presented *DiSR*, a simple yet effective framework that explicitly disentangles 3D perception from spatial reasoning through structured 3D evidence. By leveraging specialized perception models to reconstruct the physical world into reliable 3D evidence and then training an LLM with LoRA to reason solely over explicit geometric evidence, *DiSR* achieves state-of-the-art performance on multiple spatial reasoning benchmarks while requiring substantially less training data, fewer trainable parameters, and significantly lower computational cost than existing approaches. Beyond its strong empirical performance, the explicit separation of perception and reasoning improves interpretability, diagnosability, and modularity, demonstrating that disentangling perception and reasoning is a scalable and effective alternative paradigm for spatial intelligence.

AI USE STATEMENT

We used generative AI tools to polish the language of the manuscript. The authors reviewed the AI-assisted edits and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper describes the DiSR architecture, training-data construction, and evaluation setup. Appendix D details the optimization configuration, pretrained perception modules, software versions, and computational resources, with training hyperparameters summarized in Table A6. Appendices B and C provide fine-grained evaluation results and diagnostic analyses. Appendix E describes the additional task-adaptation experiment.

REFERENCES

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14455–14465. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01370. URL https://doi.org/10.1109/CVPR52733.2024.01370.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37109–37120, 2026a.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026b.

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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Yi Han, Cheng Chi, Enshen Zhou, Shanyu Rong, Jingkun An, Pengwei Wang, Zhongyuan Wang, Lu Sheng, and Shanghang Zhang. Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. *CoRR*, abs/2510.07181, 2025. doi: 10.48550/ARXIV.2510.07181. URL https://doi.org/10.48550/arXiv.2510.07181.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/413885e70482b95dcbeeddc1daf39177-Abstract-Conference.html.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/huang24ae.html.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17307–17316. IEEE, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. URL http://arxiv.org/abs/1811.00982.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023a. doi: 10.48550/ARXIV.2307.16125. URL https://doi.org/10.48550/arXiv.2307.16125.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. Hispatial: Taming hierarchical 3d spatial understanding in vision-language models. *arXiv preprint arXiv:2603.25411*, 2026.

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*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Chenyang Ma, Kai Lu, Ta Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7f2257d2b291b8d7e712c70b67e09412-Abstract-Conference.html.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan L. Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 6924–6934. IEEE, 2025a. doi: 10.1109/ICCV51701.2025.00651. URL https://doi.org/10.1109/ICCV51701.2025.00651.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan L. Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/cdd63c6c813a6ec826051b6a49f11b5e-Abstract-Conference.html.

Wufei Ma, Luoxin Ye, Celso M. de Melo, Alan L. Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 17249–17260. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.01608. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_SpatialLLM_A_Compound_3D-Informed_Design_towards_Spatially-Intelligent_Large_Multimodal_Models_CVPR_2025_paper.html.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wang25er.html.

x.ai. Realworldqa. https://x.ai/blog/grok-1.5v, 2024.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXV*, volume 15083 of *Lecture Notes in Computer Science*, pp. 131–147. Springer, 2024. doi: 10.1007/978-3-031-72698-9_8. URL https://doi.org/10.1007/978-3-031-72698-9_8.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from A single image. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp.

9009–9019. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00830. URL https://doi.org/10.1109/ICCV51070.2023.00830.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/e3a0db7c0a191854c176af1d20cdec80-Abstract-Datasets_and_Benchmarks_Track.html.

Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/8db9279f593652ee9bb2223b4a2c43fa-Abstract-Conference.html.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2899–2909. IEEE, 2023b. doi: 10.1109/ICCV51070.2023.00272. URL https://doi.org/10.1109/ICCV51070.2023.00272.

A Additional Results on Object Grounding

Table A1: **Object grounding on the 8,548-example COCO val2017 unique-instance subset.** Found is the percentage of examples for which a model returns a valid bounding box. All predictions are converted to original-image pixel coordinates before evaluation.

Model	Found ↑	mIoU ↑	IoU@0.50 ↑	IoU@0.75 ↑
Qwen2.5-VL-7B-Instruct	96.64	64.02	71.64	54.59
Qwen3-VL-8B-Instruct	89.67	71.05	78.81	66.39
SpatialReasoner	93.66	34.82	40.51	7.07
HiSpatial-3B (RGB-XYZ)	fails	–	–	–
DiSR-8B-LoRA	**99.15**	**73.99**	**81.84**	**68.17**

B Additional Results on 3DSRBench

Table A2: Fine-grained relation accuracy (%) on 3DSRBench. The best and second-best results in each row are shown in **bold** and underlined, respectively.

Category	Relation	SpatialReasoner	HiSpatial-3B	DiSR-Qwen2.5	DiSR-Qwen3
Height	Higher	51.70	**70.29**	60.43	62.75
Location	Above	**70.70**	47.71	69.36	70.38
	Closer to camera	80.10	**86.29**	76.70	80.68
	Next to	**75.50**	62.86	66.96	62.24
Orientation	In front of	63.70	**77.71**	71.22	75.29
	On the left	64.20	**85.14**	70.20	77.08
	Viewpoint	36.40	**67.43**	53.35	55.69
Multi-Object	Closer to	54.90	73.14	75.14	**77.14**
	Facing	68.20	52.57	67.34	**70.81**
	Viewpoint toward object	34.40	32.57	46.94	**50.73**
	Parallel	50.10	47.43	**62.54**	57.82
	Same direction	55.50	50.29	56.10	**59.59**

Fine-grained Relation Analysis. Table A2 presents fine-grained comparisons across the 12 relation types in 3DSRBench. HiSpatial achieves strong performance on relations involving height, camera depth, and single-object orientation, including *Higher*, *Closer to camera*, *In front of*, *On the left*, and *Viewpoint*, benefiting from its large-scale spatial training. In contrast, *DiSR* consistently excels at multi-object spatial relations, achieving the best performance on all five such categories. Specifically, DiSR-Qwen3 leads on *Closer to*, *Facing*, *Viewpoint toward object*, and *Same direction*, while DiSR-Qwen2.5 performs best on *Parallel*. These relations require aggregating geometric information across multiple objects or reasoning in object-dependent coordinate systems, which is challenging to learn solely from images. The superior performance of *DiSR* on these tasks highlights the advantage of explicit 3D evidence for reliable and compositional spatial reasoning.

C Shortcut and Error Analysis on CV-Bench-3D

Although CV-Bench-3D is designed to evaluate spatial reasoning beyond two-dimensional image-plane cues, its Distance and Depth subsets still contain strong correlations between simple 2D box geometry and the ground-truth answers. To better understand these effects, we first analyze task-specific 2D heuristics for each subset and then exploit the explicit intermediate representations in *DiSR* to identify the remaining sources of error.

Distance: Image-Plane Proximity

The Distance subset evaluates whether a model can determine which of two candidate objects is closer to a reference object in 3D space. However, in many samples, the object that is closer to the reference object in 3D also exhibits a smaller 2D center distance. As a result, models may

achieve high accuracy by exploiting 2D spatial correlations rather than recovering the underlying 3D relationship, making the reported performance on this subset an imperfect measure of true 3D reasoning ability.

2D Proximity Heuristic. Following the analysis of SpatialReasoner, we consider a simple heuristic that uses neither depth nor reconstructed 3D geometry. Let $\mathbf{c}_r$ denote the 2D center of the red reference box, and let $\mathbf{c}_b$ and $\mathbf{c}_g$ denote the centers of the blue and green candidate boxes, respectively. The heuristic selects the candidate with the smaller image-plane distance to the reference:

$$\hat{y}_{\mathrm{2D}} = \arg\min_{o \in \{b,g\}} \|\mathbf{c}_r - \mathbf{c}_o\|_2 . \tag{9}$$

Despite its simplicity, this rule obtains **80.17%** accuracy on the 600 Distance questions of CV-Bench-3D. In contrast, it obtains only **34.30%** on the *"Multi-Object Closer to"* relation of 3DSR-Bench, where image-plane proximity is substantially less predictive of the annotated 3D relationship.

2D-Consistent and 2D-Conflict Subsets. To separate samples where image-plane cues correlate with the underlying 3D relation from those where they fail, we partition the 600 Distance questions into two disjoint subsets. The *2D-consistent* subset consists of 481 samples where the 2D heuristic prediction, $\hat{y}_{\mathrm{2D}}$, agrees with the ground-truth answer, whereas the remaining 119 samples form the *2D-conflict* subset where the heuristic produces the opposite prediction. We also report the accuracy for the two subsets.

Table A3: Analysis of distance-related reasoning on CV-Bench-3D and 3DSRBench. The CV-Bench-3D Distance subset is divided according to whether the 2D bounding-box-center heuristic agrees with the ground-truth answer. The final column reports accuracy on the corresponding *Multi-Object Closer to* relation of 3DSRBench.

Method	CV-Bench-3D Distance			3DSRBench
	Overall	2D-Consistent (481)	2D-Conflict (119)	Multi-Object Closer to
2D Heuristic	80.17	100.00	0.00	34.30
SpatialReasoner	72.83	75.26	63.03	54.90
Qwen3-VL-8B-Instruct	89.83	98.96	52.94	61.40
DiSR-8B-LoRA	**91.67**	95.84	**74.79**	**77.14**

Discussion on Results. Table A3 jointly reveals a cross-benchmark discrepancy and substantial shortcut dependence within CV-Bench-3D Distance. The 2D heuristic achieves 80.17% accuracy on CV-Bench-3D but only 34.30% on the corresponding 3DSRBench, indicating that image-plane proximity is substantially more predictive on CV-Bench-3D. Qwen3-VL-8B-Instruct similarly drops from 98.96% on the 2D-consistent subset to 52.94% on the 2D-conflict subset. In contrast, DiSR-8B-LoRA improves conflict accuracy to 74.79% and consistently achieves the strongest result, 77.14%, on the corresponding 3DSRBench. These results indicate that reasoning over explicit 3D evidence is a more reliable alternative for spatial intelligence.

Figure A3 provides qualitative examples that complement the quantitative analysis. In the CV-Bench-3D example, the correct answer can be inferred almost directly from the relative positions of the highlighted 2D bounding boxes, indicating the presence of strong image-plane cues. In contrast, applying the same 2D heuristic to the 3DSRBench example leads to an incorrect prediction, despite all queried objects being clearly visible. This discrepancy highlights the difference in shortcut availability across benchmarks and motivates our further analysis of 2D shortcut dependence in CV-Bench-3D.

DEPTH: BOTTOM-POSITION SHORTCUT

The Depth subset evaluates whether a model can determine which of two highlighted objects is closer to the camera. Although this task requires camera-depth reasoning, the annotated depth relation is strongly correlated with the vertical positions of the object boxes. In many ground-plane scenes, the object closer to the camera is projected lower in the image, creating a potentially strong image-plane shortcut.

CVBench-3D (Distance)



Question: Estimate the real-world distances between objects in this image. Which object is closer to the monitor (highlighted by a red box), the chair (highlighted by a blue box) or the keyboard (highlighted by a green box)?
Answer: Keyboard.

3DSRBench (Multi-Object Closer to)
(bounding box added for visualization)



Question: Consider the real-world 3D locations of the objects.Which is closer to the policeman, the camera or the yellow taxi?
Answer: Camera.

Figure A3: Qualitative examples illustrating the 2D shortcut in CV-Bench-3D Distance and its failure to transfer reliably to 3DSRBench. **Left:** In the CV-Bench-3D Distance example, the keyboard occupies an image-plane region close to and partially overlapping the red reference box around the monitor, whereas the chair is projected substantially farther away. The 2D proximity heuristic therefore selects the keyboard, matching the annotated answer as well as the predictions of Qwen3-VL-8B-Instruct and DiSR-8B-LoRA. **Right:** In the 3DSRBench *Multi-Object Closer to* example, the center of the yellow-taxi box lies closer to the police-officer box in the image plane than the center of the camera box. Nevertheless, the annotated 3D answer is the camera, demonstrating that image-plane proximity is not a reliable proxy for the underlying 3D relation. Bounding boxes in the 3DSRBench image are added only for visualization and are not part of the original benchmark input.

2D Bottom-Position Heuristic. Let y_r^{bot} and y_b^{bot} denote the bottom coordinates of the red and blue object boxes, respectively, where the image y-coordinate increases downward. We define a simple 2D heuristic that selects the object with the larger bounding-box-bottom coordinate, corresponding to a lower position in the image:

$$\hat{y}_{\text{bottom}} = \arg\max_{o \in \{r,b\}} y_o^{\text{bot}}. \tag{10}$$

Despite using neither depth information nor reconstructed 3D geometry, this heuristic correctly answers 512 of the 600 Depth questions, achieving **85.33%** accuracy. We therefore partition the samples into 512 *2D-consistent* cases, for which the heuristic agrees with the ground-truth answer, and 88 *2D-conflict* cases, for which it selects the incorrect object.

Figure A4 provides representative examples from the two subsets. In the *2D-consistent* example, the table has a lower bounding-box-bottom position than the bookcase and is also annotated as the object closer to the camera. The bottom-position heuristic therefore produces the correct answer without using explicit depth information. In the *2D-conflict* example, the chair has a lower bounding-box-bottom position than the hanging lamp, causing the same heuristic to select the chair, whereas the annotated closer object is the lamp. These examples illustrate both the strong predictive correlation captured by the heuristic and the cases in which this image-plane cue conflicts with the underlying camera-depth relation.

Discussion on Results. Table A4 confirms that CV-Bench-3D Depth contains a strong bottom-position shortcut. The simple 2D heuristic correctly answers 512 of the 600 questions, indicating that the annotated closer object is usually projected lower in the image. Figure A4 qualitatively

Table A4: Accuracy (%) on the CV-Bench-3D Depth subset after dividing the samples according to whether the 2D bounding-box bottom-position heuristic agrees with the ground-truth answer. The heuristic obtains 100.00% on the 2D-consistent subset and 0.00% on the 2D-conflict subset by construction. The best result among the learned models in each column is shown in **bold**.

Method	Overall	2D-Consistent (512)	2D-Conflict (88)
2D Heuristic	85.33	100.00	0.00
SpatialReasoner	87.00	87.30	85.23
Qwen3-VL-8B-Instruct	**95.50**	**96.09**	**92.05**
DiSR-8B-LoRA	92.83	93.55	88.64

illustrates both sides of this correlation: the cue directly yields the correct answer in a 2D-consistent sample but becomes misleading in a 2D-conflict sample.

Qwen3-VL-8B-Instruct and DiSR-8B-LoRA retain 92.05% and 88.64% accuracy, respectively, on the 2D-conflict subset, showing that both models can often recover the correct camera-depth relation even when the bottom-position cue is misleading. Moreover, Qwen3-VL-8B-Instruct outperforms DiSR-8B-LoRA on both the 2D-consistent and 2D-conflict subsets. Its overall advantage therefore cannot be explained solely by stronger exploitation of the bottom-position shortcut. The object-size analysis in Table A5 instead shows that the remaining errors of DiSR-8B-LoRA are concentrated in samples containing small target objects, for which automatically constructed object-level 3D evidence is less reliable.

Table A5: Depth accuracy (%) stratified by the image-plane size of the ground-truth answer object. The 600 samples are divided into five equally sized groups of 120 samples. "DiSR w/ GT Object Grounding" uses ground-truth object regions while retaining automatically constructed 3D evidence. "DiSR w/ GT 3D Evidence" additionally replaces the automatically constructed evidence with ground-truth 3D evidence, achieving 100% accuracy in every size group.

Answer-object size	Median box area (%)	Qwen3-VL-8B Instruct	DiSR-8B LoRA	DiSR w/ GT Object Grounding	DiSR w/ GT 3D Evidence
Smallest 20%	0.236	89.17	84.17	85.00	100.00
Small	0.786	97.50	95.00	95.00	100.00
Medium	2.184	95.83	92.50	93.33	100.00
Large	5.618	96.67	94.17	95.00	100.00
Largest 20%	14.453	98.33	98.33	98.33	100.00

Structured 3D Evidence for Small Objects

Table 5 in the main paper identifies that the performance bottleneck of *DiSR* on CV-Bench-3D lies in the inaccurate 3D evidence. As we observe that the images in CV-Bench-3D are often in low-resolution, we further investigate whether structured 3D evidence will become less reliable when the queried object occupies only a small image region. For each question, we measure the image-plane size of the ground-truth answer object using its annotated 2D bounding box:

$$s_{\text{ans}} = \frac{w_{\text{ans}} h_{\text{ans}}}{WH}, \tag{11}$$

where W and H denote the image width and height, respectively. The 600 Depth samples are sorted according to s_{ans} and divided into five equally sized groups of 120 samples, ranging from the smallest to the largest answer objects.

Discussion on Results. Table A5 shows that the lower performance of *DiSR* is caused by samples with small target objects. For the smallest-object group, where the objects occupy only **0.236%** of the image area, Qwen3-VL-8B-Instruct outperforms DiSR-8B-LoRA by 5.00%, while the 3D-evidence construction error rate reaches 15.00%. In contrast, for the largest-object group, where objects occupy 14.453% of the image area, *DiSR* achieves comparable accuracy to Qwen3-VL-8B-Instruct, with the 3D-evidence construction error rate reduced to only 1.67%. Figure A5 presents four representative failure cases involving small image-plane target objects. Across these examples,

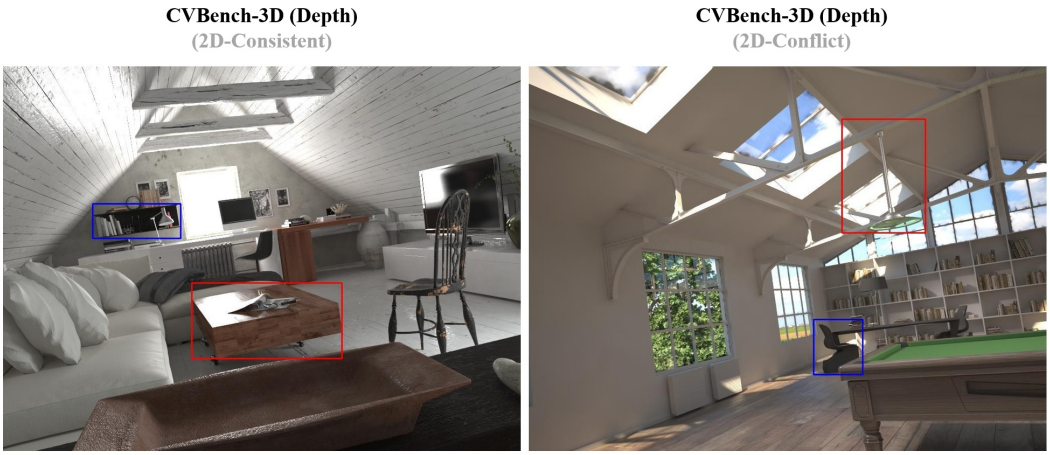


Figure A4: Representative examples from the 2D-consistent and 2D-conflict subsets of CV-Bench-3D Depth analyzed in Table A4. **Left:** In the 2D-consistent example, the table has a lower bounding-box-bottom position than the bookcase and is also annotated as the object closer to the camera. The bottom-position heuristic therefore produces the correct answer without using explicit depth information. **Right:** In the 2D-conflict example, the chair has a lower bounding-box-bottom position than the hanging lamp, causing the same heuristic to select the chair, whereas the annotated closer object is the lamp. These examples illustrate when the bottom-position shortcut agrees with the underlying camera-depth relation and when it becomes misleading.

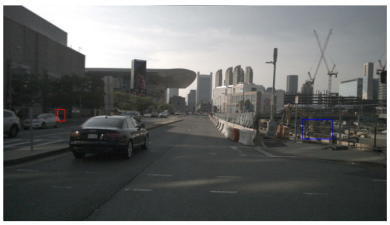
the queried objects occupy only a limited image region, making object-level depth and geometric estimation particularly challenging. All four examples are selected from samples that remain incorrect under GT Object Grounding but are recovered with GT 3D Evidence. This pattern indicates that the remaining errors arise from inaccurate object-level geometry rather than incorrect object references. These examples qualitatively support the size-conditioned trend reported in Table A5.

D Reproducibility Details

Optimization. We use a QLoRA-style setup in both stages. The frozen backbone is loaded with 4-bit NF4 quantization, double quantization, and bfloat16 compute, while only the LoRA parameters are optimized. LoRA is applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, with rank 128, scaling factor 256, dropout 0.05, and no trainable bias. We use `paged_adamw_8bit` with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, zero weight decay, and a maximum gradient norm of 1.0. The learning rate follows a cosine schedule after a 3% linear warmup. Training uses bfloat16, gradient checkpointing, and a single GPU. The random seed is set to 42.

Perception Modules. All expert perception models remain frozen during training and evaluation. The stack uses SAM 2.1 HQ with a Hiera-L backbone for object masks, Metric3D v2 with a ViT-Giant backbone for metric depth, Perspective Fields with the ParamNet-360Cities checkpoint and WildCamera with its all-data checkpoint for camera geometry, and Orient Anything with a DINOv2-L backbone for object orientation. No expert model is fine-tuned on the evaluation benchmarks.

Software and Compute. DiSR-8B is trained with Python 3.10.20, PyTorch 2.6.0+cu118, Transformers 4.57.6, PEFT 0.19.1, Accelerate 1.14.0, and bitsandbytes 0.49.2 on a single NVIDIA RTX 4080 SUPER under Linux kernel 5.15.0-78-generic. Stages 1 and 2 require 19.50 and 39.25 GPU-hours, respectively, for a total of 58.75 GPU-hours.



Question: Which object is closer to the camera taking this photo, the pedestrian (highlighted by a red box) or the truck (highlighted by a blue box)?
Gold answer: Pedestrian.
Qwen3-VL-8B-Instruct: Pedestrian. **DiSR-8B-LoRA:** Truck.



Question: Which object is closer to the camera taking this photo, the traffic cone (highlighted by a red box) or the pedestrian (highlighted by a blue box)?
Gold answer: Traffic cone.
Qwen3-VL-8B-Instruct: Traffic cone. **DiSR-8B-LoRA:** Pedestrian.



Question: Which object is closer to the camera taking this photo, the traffic cone (highlighted by a red box) or the truck (highlighted by a blue box)?
Gold answer: Traffic cone.
Qwen3-VL-8B-Instruct: Traffic cone. **DiSR-8B-LoRA:** Truck.



Question: Which object is closer to the camera taking this photo, the truck (highlighted by a red box) or the pedestrian (highlighted by a blue box)?
Gold answer: Truck.
Qwen3-VL-8B-Instruct: Truck. **DiSR-8B-LoRA:** Pedestrian.

Figure A5: Four representative failure cases involving small image-plane target objects from CV-Bench-3D Depth, corresponding to the size-stratified analysis in Table A5. Across these examples, Qwen3-VL-8B-Instruct produces the correct answer, whereas DiSR-8B-LoRA makes an incorrect prediction. All four examples are selected from samples that remain incorrect under GT Object Grounding but are recovered with GT 3D Evidence, indicating errors in the automatically constructed object-level geometry.

Table A6: Training hyperparameters for the reported DiSR-8B model.

Setting	DiSR-8B
Stage-1 training instances	112,157
Stage-2 training instances	216,000
Epochs per stage	1
Maximum sequence length	1,024
Per-device batch size	1
Gradient accumulation	16
Effective batch size	16
Stage-1 learning rate	2×10^{-5}
Stage-2 learning rate	2×10^{-5}
Stage-1 optimizer steps	7,010
Stage-2 optimizer steps	13,500
Warmup ratio	0.03
Stage-1 warmup steps	211
Stage-2 warmup steps	405
Optimizer	`paged_adamw_8bit`
Learning-rate scheduler	cosine
Weight decay	0
Maximum gradient norm	1.0
LoRA rank / scaling	128 / 256
LoRA dropout	0.05
Random seed	42

E Error Analysis on SPAR-Bench

The main paper shows that the original training distribution provides limited supervision for relational-selection tasks in SPAR-Bench. To investigate whether targeted supervision can alleviate this limitation, we conduct an additional post-freeze diagnostic experiment. Starting from the frozen Epoch-2 *DiSR-8B-LoRA* model, we construct a 2,816-example adaptation set, with half of the examples covering two previously unseen SPAR-Bench task types and the other half replaying the original spatial reasoning data to mitigate forgetting. We train a new LoRA adapter for 10 epochs and denote the resulting model as *DiSR-8B-LoRA**. We evaluate this model using the same full-chain protocol and report the results separately.

Results Discussion. As shown in Table 2, *DiSR-8B-LoRA** improves over *DiSR-8B-LoRA* by **11.23%** on SPAR-Bench, demonstrating that task-specific supervision effectively addresses the identified limitation. Meanwhile, Table A7 presents a fine-grained comparison on 3DSRBench, where *DiSR-8B-LoRA** maintains comparable overall performance with *DiSR-8B-LoRA*, indicating that the additional task-focused adaptation improves SPAR-Bench performance without sacrificing general spatial reasoning ability.

Table A7: Fine-grained relation accuracy (%) on 3DSRBench, including the task-adapted model. The best and second-best results in each row are shown in **bold** and underlined, respectively.

Category	Relation	SpatialReasoner	HiSpatial-3B	DiSR-Qwen2.5	DiSR-8B-LoRA	DiSR-8B-LoRA*
Height	Higher	51.70	**70.29**	60.43	62.75	62.61
Location	Above	**70.70**	47.71	69.36	70.38	63.29
	Closer to camera	80.10	**86.29**	76.70	80.68	80.68
	Next to	**75.50**	62.86	66.96	62.24	61.36
Orientation	In front of	63.70	**77.71**	71.22	75.29	73.26
	On the left	64.20	**85.14**	70.20	77.08	78.51
	Viewpoint	36.40	**67.43**	53.35	55.69	51.60
Multi-Object	Closer to	54.90	73.14	75.14	77.14	**78.57**
	Facing	68.20	52.57	67.34	**70.81**	67.63
	Viewpoint toward object	34.40	32.57	46.94	**50.73**	49.27
	Parallel	50.10	47.43	**62.54**	57.82	54.28
	Same direction	55.50	50.29	56.10	**59.59**	59.30