

MV-STRIDE: Enabling MLLMs to Master Multi-View Spatial Reasoning via Hierarchical Capability Modeling

Jin Xu¹, Xiaojian Huang¹, Zhuodong Luo¹, Zhihong Zhang¹, Xin Liu²,
Jiansheng Wei², Xinzhi Wang², Jie Zhao³, and Xuejin Chen¹*

¹ MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China, China

² Huawei Noah's Ark Lab

³ Hefei University of Technology, China

Abstract. Despite the rapid progress of Multimodal Large Language Models (MLLMs) in 2D vision-language tasks, robust multi-view spatial reasoning remains a fundamental bottleneck due to the lack of structured 3D cognitive pathways in existing datasets. To address this, we introduce MV-STRIDE, a **M**ulti-**V**iew hierarchical **S**pa**T**ial **R**easoning dataset with **I**nterdependent and **D**ecomposed capabilities. Moving beyond flat data structures, MV-STRIDE explicitly models the dependency relationships between foundational perception, scene understanding, and complex contextual reasoning, providing a coherent learning pathway aligned with human spatial cognition. We develop a systematic QA generation pipeline leveraging diverse 3D scene sources that enforces cross-view dependency constraints to prevent single-view solvability, generating multi-level spatial reasoning tasks supported by cognitively grounded chain-of-thought supervision for complex inference. Extensive evaluations demonstrate that our multi-stage training framework based on our hierarchical dataset achieves state-of-the-art performance across multiple spatial reasoning benchmarks, notably the multi-view oriented MMSI-Bench. Our approach enables MLLMs to maintain robust, 3D-consistent spatial reasoning across diverse viewpoints. The code and dataset are available at <https://co1dspring.github.io/MV-STRIDE/>.

Keywords: Multimodal Large Language Model · Dataset Construction
· Multi-view Spatial Reasoning

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) [16, 17, 26, 29, 30] have excelled in general vision-language tasks, especially fundamental spatial perception, which are essential for downstream applications such as embodied intelligence and autonomous driving. Despite these advances, robust spatial reasoning under multi-view visual conditions remains major unresolved challenges.

* Corresponding authors.

Extensive evaluations of multi-view spatial reasoning benchmarks [14, 22, 27, 43, 45] consistently reveal a significant performance gap between humans and MLLMs. Humans naturally integrate shared objects or structures across views into a coherent 3D cognitive map enabling flexible viewpoint transformations and reasoning about arbitrary spatial relationships. In contrast, current MLLMs struggle to reliably establish such correspondences, maintain geometric consistency across views, or perform stable viewpoint-dependent reasoning. Error analyses of existing benchmarks [22, 43, 45] indicate that failures predominantly arise from deficiencies in scene-level reconstruction, cross-view correspondence, and viewpoint transformation, rather than from isolated perceptual errors.

A key cause of multi-view bottleneck lies in the design limitations of existing spatial reasoning datasets. Most spatial reasoning datasets [5, 6, 8, 20, 24, 25, 36, 39] focus on single-view 2D planes, failing to support the formation of consistent 3D spatial cognition. Although recent efforts have introduced multi-view data [23, 32, 40, 42, 44, 47, 48], many datasets emphasize low-level quantitative perception (*e.g.* depth estimation), which strengthens isolated perceptual skills but does not support the development of structured 3D reasoning. Furthermore, multi-view spatial reasoning is frequently treated as a flat or loosely defined capability, with tasks designed independently and without explicit dependency between perception, scene understanding, and reasoning. As a result, models lack a clear learning pathway from basic spatial perception to high-level multi-view reasoning, hindering their ability to generalize to complex, multi-step spatial inference.

To bridge this gap, we introduce MV-STRIDE, a **M**ulti-**V**iew hierarchical **S**pa**T**ial **R**easoning dataset with **I**nterdependent and **D**ecomposed capabilities. Aligned with human cognition, MV-STRIDE organizes tasks into three levels: single-view perception, cross-view scene understanding, and complex contextual spatial reasoning. Unlike simple difficulty-based hierarchies, our framework explicitly models the dependency relationships among different spatial capabilities across levels. This design provides a coherent pathway for MLLMs to acquire stable 3D spatial awareness and reasoning behavior aligned with human cognition.

Following this design, we develop a pipeline that leverages diverse 3D-grounded assets to generate MV-STRIDE. Our pipeline employs a top-down generation strategy that simultaneously populates all three levels of the hierarchy. We first construct complex Level III tasks—specifically enforced by cross-view dependency constraints to preclude single-view shortcuts—and then decompose them into prerequisite Level I and II sub-questions. These tasks are organized into spatial cognition problem groups, where the ordered answers to sub-questions naturally form the constituent steps of Chain-of-Thought (CoT) annotations. This structured data facilitates a capability-aligned training strategy, spanning from supervised fine-tuning to reinforcement learning, ensuring model reasoning remains strictly aligned with the human spatial cognition process. Models trained with our dataset and methodology achieve state-of-the-art performance

across multiple spatial reasoning benchmarks, particularly in multi-view settings, thereby validating the superior effectiveness of our proposed approach.

Our main contributions are summarized as follows:

- We introduce a hierarchical, capability-dependent organization of multi-view spatial reasoning tasks, which explicitly models the dependency structure from foundational spatial perception to high-level multi-view contextual reasoning.
- We design a cognition-driven QA generation pipeline and construct a multi-level multi-view spatial reasoning dataset MV-STRIDE with accurate 3D supervision, where tasks across different levels are explicitly connected through capability dependencies and organized into coherent spatial cognition problem groups.
- We empirically validate the effectiveness of the proposed dataset by multi-stage training, demonstrating consistent improvements on multiple spatial reasoning benchmarks.

2 Related Work

2.1 Multimodal Spatial Reasoning Dataset

Recently, a plethora of large-scale spatial reasoning datasets has emerged to bolster the 3D awareness of MLLMs. Early efforts relied on pseudo-labeling pipelines to automatically generate extensive training annotations. SpatialVLM [6] pioneered automated 3D annotation for internet images, followed by refinements in region-level scene graphs [8] and RGB-D perception [5]. To mitigate the inherent inaccuracies of pseudo-labels, recent research has shifted toward precise 3D ground-truth. Datasets such as SPAR-7M [48], RefSpatial [49], SAT [32], VST [44] and CA-1M [10] leverage real-world reconstructions (e.g., ScanNet++ [46]) or simulated environments [12] as data sources to provide accurate supervision across various spatial reasoning tasks. While these works have scaled single-view reasoning, they lack the structured, multi-layered pathways necessary to systematically foster complex 3D geometric consistency.

2.2 Multimodal Multi-view Spatial Reasoning Dataset

Despite the growth in data scale, robust multi-view spatial reasoning remains a bottleneck as many existing datasets either treat multi-frame inputs as isolated contexts [10] or provide limited multi-view task categories [32, 44, 48]. Dedicated multi-view datasets like MultiSPA [42] offer large-scale perception labels (e.g., depth, motion) but are restricted to low-level quantitative estimation. Other efforts focus on isolated capabilities, such as viewpoint recognition [47], point correspondence [40], or grounding-based reasoning [23]. Unlike these narrowly focused or perception-heavy datasets, our MV-STRIDE introduces a comprehensive, three-level hierarchy that explicitly models the dependency between foundational perception and high-level contextual reasoning, ensuring robust 3D-consistent spatial inference.

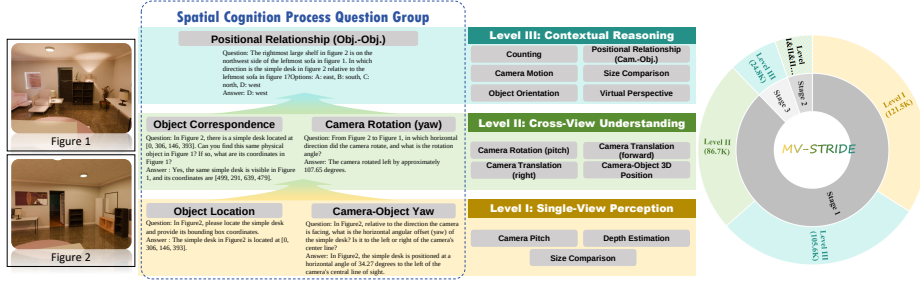


Fig. 1: Overview of our MV-STRIDE. It encompasses a hierarchy of spatial tasks across three levels and 18 categories, sourced from over 600 synthetic and real indoor environments. This cognition-aligned structure supports the entire training lifecycle from Supervised Fine-Tuning (SFT) to Reinforcement Learning (RL), enabling MLLMs to develop human-like, multi-view consistent spatial reasoning capabilities.

3 Method

To address the bottleneck of systematic multi-view spatial reasoning capability in existing MLLMs, we propose a hierarchical, cognition-driven dataset construction methodology tailored for 3D-consistent multi-view spatial reasoning. Our method is grounded in a key observation: robust spatial reasoning does not emerge from isolated tasks, but from a progressive acquisition of spatial perception, scene-level understanding, and high-level contextual reasoning abilities. Consequently, we organize spatial reasoning tasks following a "Perception, Modeling and Reasoning" cognitive hierarchy and construct a layered question-answering (QA) dataset that explicitly models capability dependencies across levels.

In the following, Sec. 3.1 introduces the hierarchical definition of spatial reasoning tasks and their capability dependencies. Sec. 3.2 describes the data generation pipeline used to construct the multi-level QA dataset. Sec. 3.3 presents our multi-stage training strategy built upon the proposed dataset.

3.1 Hierarchical Definition of Spatial Reasoning Capabilities

We decompose multi-view spatial reasoning into three tightly coupled levels, each corresponding to a distinct stage in human spatial cognition and addressing a specific limitation of current MLLMs. These levels jointly form a unified spatial reasoning training system. The hierarchical architecture and task composition of MV-STRIDE are illustrated in Fig. 1.

Level I: Single-View Spatial Perception. Level I focuses on the establishment of robust, egocentric spatial perception within a single perspective. As the bedrock of the MV-STRIDE hierarchy, this level transitions the model's focus to quantitative geometric awareness. Specifically, the task types include recognition of the spatial attributes of objects and cameras, and relative position

judgment. By ensuring precise perception at the single-view level, we prevent hallucinated geometry from propagating into subsequent 3D scene understanding stages. From a cognitive perspective, it mimics the human ability to instantly perceive the layout of a visible environment, providing the necessary spatial clues to build a global scene representation.

Level II: 3D-Consistent Scene Understanding. Level II focuses on scene-level spatial understanding by establishing geometric correspondences and view-point transformations across two perspectives. The primary objective is to bridge isolated perspectives into a coherent 3D cognitive map. Tasks at this level require MLLM to resolve inter-view relationships, including cross-view object correspondence and relative camera motion estimation (covering translation and rotation). By mastering these capabilities, the model is encouraged to transition from independent image processing to a global modeling of the scene structure, effectively overcoming spatial inconsistency and providing the necessary structural foundation for subsequent complex reasoning.

Level III: Multi-View High-Level Contextual Reasoning. Level III focuses on high-order, context-dependent spatial reasoning, representing the most comprehensive task suite for multi-view spatial reasoning to date. Building upon the structural representations established in Level II, this level requires the model to perform sophisticated inference across a diverse set of tasks, including multi-view object counting, measurement estimation, object orientation reasoning and positional relationship reasoning across multiple reference frames. To achieve high-level proficiency, these tasks are designed to necessitate virtual viewpoint transformations and allocentric reference frame reasoning based on full 6-DoF camera states, pushing the model’s complexity well beyond simple ego-centric observation.

Crucially, to prevent the model from relying on single-view shortcuts, we enforce a strict cross-view dependency constraint: every question is designed to be unsolvable from any single perspective. For example, the model may be asked to determine the spatial direction of an object visible only in View B relative to the camera’s position in View A. This design encourages the model to perform accurate perception and comprehensive cross-view information integration before reasoning, guiding it to develop genuine 3D-aware spatial reasoning ability instead of merely reading 2D information.

Spatial Cognition Process Question Group. Based on the hierarchical definition of spatial reasoning tasks, we further introduce a cognition-process-based grouping strategy to explicitly model the reasoning trajectory required for complex tasks. In this approach, we treat the high-level multi-view contextual reasoning problems in Level III as the apex. For each complex problem, we perform a top-down decomposition of the underlying spatial cognitive requirements, reverse-engineering the necessary perceptual and scene-understanding

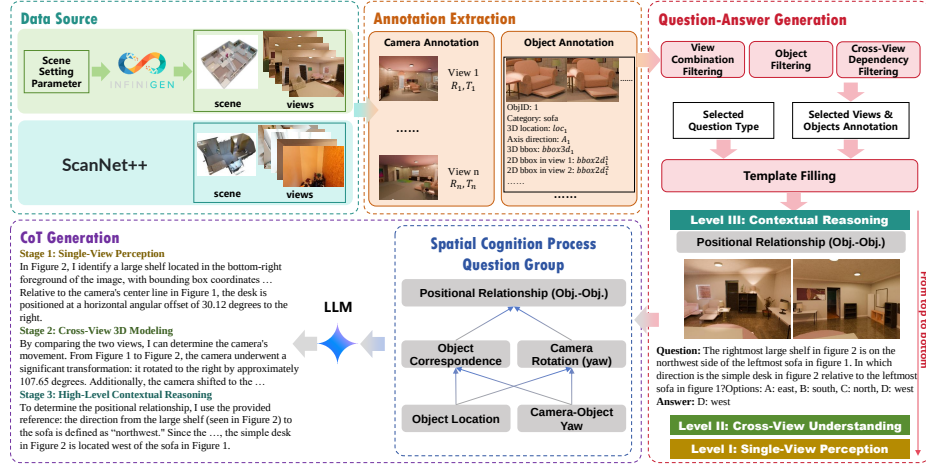


Fig. 2: Overview of MV-STRIDE construction pipeline. We convert 3D scene assets into hierarchical multi-view spatial reasoning tasks. Scene annotations are first extracted from synthetic and real-world scene datasets. A filtering stage selects valid cameras and objects while enforcing cross-view dependency constraints. Based on the filtered metadata, hierarchical QA pairs are generated across three spatial cognition levels. Finally, cognition-process-based question groups are used to synthesize chain-of-thought reasoning annotations for high-level tasks.

steps. These intermediate capabilities are then matched and instantiated as specific sub-questions within Level I and Level II. Consequently, for a given scene and viewpoint set, questions across all three levels are organized into a coherent spatial cognition process question group.

By explicitly encoding these capability dependencies and logical transitions, this design ensures that high-level reasoning is strictly grounded in verifiable perceptual and scene-level understanding, thereby guiding models to follow a cognitively plausible reasoning trajectory that transitions from local perception to global integration. Furthermore, this structured organization allows for the construction of principled, long-chain reasoning supervision, providing an efficient pathway to acquire stable 3D spatial awareness and robust multi-view reasoning behavior.

3.2 Data Construction Pipeline

We develop a systematic, cognition-driven pipeline to transform raw 3D scene data into structured, hierarchical spatial reasoning tasks, as illustrated in Fig. 2. By grounding natural-language reasoning in verifiable 3D metadata, this framework bridges the gap between low-level visual perception and high-level contextual inference.

Scene Assets and Preprocessing. Our pipeline leverages two distinct types of data sources to ensure both realism and structural diversity: procedural 3D scene generators and real-world 3D reconstruction datasets.

Procedural 3D Scene Generation. We utilize the Infinigen [31] framework to procedurally generate a vast array of novel indoor environments. This allows us to create diverse room layouts, varied object distributions, and randomized viewpoint trajectories. For each generated scene, we render multi-view images while simultaneously capturing exhaustive ground-truth metadata, including 3D object bounding boxes, precise camera intrinsics/extrinsics, and view-dependent 2D instance segmentations.

Real-World Reconstruction. To ground our dataset in real-world complexity, we incorporate ScanNet++ [46], which provides high-fidelity 3D reconstructions of actual indoor spaces. From this source, we directly extract the requisite geometric and semantic information, ensuring that our model encounters realistic textures, lighting, and spatial configurations.

By integrating these dual sources, we establish a robust foundation of accurate 3D object and camera annotations. This geometric grounding is essential for maintaining strict multi-view consistency and generating high-precision QA pairs across our hierarchical framework.

Viewpoint and Object Filtering. To ensure the dataset effectively facilitates 3D-aware learning, we implement a rigorous multi-stage filtering process:

Camera Pair Selection. We filter camera combinations based on their spatial proximity and orientation. Pairs with excessively large or negligible spatial separation are excluded to maintain a reasonable level of difficulty. Crucially, we require each camera pair to share a minimum number of common objects, providing sufficient spatial anchors to ensure that cross-view relationships are solvable for observers.

Object Visibility and Ambiguity Control. We select target objects that are clearly visible and can be uniquely identified through unambiguous text descriptions. Objects that are heavily occluded or too small to provide reliable visual cues are discarded.

Cross-View Dependency Constraint. To uphold the single-view unsolvability of Level III tasks, we apply deterministic rules during selection. Specifically, we only retain questions that require integrating information across the entire view set, such as referencing an object visible only in View A from the camera coordinate system of View B, or evaluating spatial relationships between multiple objects that never co-occur in a single frame. This eliminates shortcuts based on local cues, compelling the model to synthesize a globally consistent scene representation to arrive at the correct answer.

To empirically validate this constraint, we randomly sample 1K Level III questions and evaluate Gemini-3-Flash under two input settings. When provided with only a random single view, the model obtains 31.5% accuracy, while using the full multi-view inputs improves the accuracy to 41.2%. This gap indicates

that Level III questions require cross-view information integration and cannot be reliably solved from a single-view shortcut.

Hierarchical QA Generation. Following the cognition-process-based grouping strategy, the generation proceeds in a top-down manner. We first generate complex Level III questions based on the filtered viewpoints and selected objects. Once a Level III task is defined, the pipeline automatically identifies the requisite perceptual and scene understanding capabilities and generates the corresponding sub-questions of Level I and Level II. To bridge the gap between geometric data and natural language, our pipeline first transforms the raw 3D metadata of filtered objects and viewpoints into structured spatial semantic information, which is then used to populate our predefined QA templates. To ensure high response quality and linguistic diversity, we utilize human-designed templates enriched by Large Language Models to achieve rich variability. While Level I and Level II tasks are formulated as short-answer questions to foster perceptual precision, Level III tasks are structured as multiple-choice questions to facilitate the learning of complex, multi-step spatial inference. This hierarchical design ensures that all tasks, ranging from low-level perception to high-level reasoning, are derived from a unified 3D scene representation and the same set of multi-view visual inputs, maintaining strict cross-level geometric consistency.

Construction of Long-Chain Reasoning Annotations. Leveraging the cognition-process-based question groups, we construct high-quality CoT annotations for Level III tasks. For each high-level problem at the group’s apex, we treat its associated Level I and Level II sub-questions and their ground-truth answers as necessary prerequisites. These components are sequentially organized and fed into an LLM [15], which serves as a linguistic synthesizer to refine them into a coherent natural-language reasoning chain.

This strategy ensures that each CoT follows the three essential stages of human spatial cognition: local perception, cross-view integration, and high-level contextual inference. By grounding the final answer in verifiable intermediate steps, our approach provides transparent and interpretable supervision for long-chain reasoning. This enables the scalable generation of complex spatial reasoning chains without manual annotation, that remains strictly aligned with the underlying cognition logic, as further detailed in the Appendix.

Quality Verification. We further conduct sampled quality verification through human and LLM-based checks for the generated QA pairs and CoT annotations, yielding an overall accuracy of 88.0% for QA pairs and 97.5% for CoT annotations. These results indicate that MV-STRIDE maintains reliable QA and CoT quality at scale, with only minor noise in automatically generated samples. More details of the verification protocol and error analysis are provided in the Appendix.

3.3 Multi-Stage Training Strategy

To fully exploit the hierarchical and grouped nature of our dataset, we design a progressive three-stage training strategy that systematically develops the model’s spatial intelligence from fundamental perception to complex reasoning.

Stage 1: Foundational Spatial SFT. In the first stage, we perform SFT on the base model using 60% of the scene data, encompassing tasks from Level I to Level III. This stage is designed to equip the model with foundational capabilities in spatial perception, cross-view scene understanding, and multi-view contextual reasoning.

Stage 2: Cold-Start with CoT. While the first stage establishes core capabilities, it does not inherently enable the model to generate interpretable reasoning chains. Therefore, we utilize 20% of the scenes from ScanNet++ to construct CoT annotations for cold-start training. This stage focuses on teaching the model the structured reasoning templates required for complex multi-view tasks, fostering an initial reasoning awareness.

Stage 3: Multi-view Spatial Reasoning RL. In the final stage, we employ the Group Relative Policy Optimization (GRPO) [33] algorithm to further refine the model. We first apply the model trained at Stage 1 to perform multiple inference passes on the remaining data and select a subset of samples with moderate difficulty based on empirical correctness statistics. Using these samples, the model then performs reinforcement learning to further explore diverse reasoning paths, building upon the reasoning templates acquired during the cold-start stage. This reinforcement learning process encourages the model to find an optimal balance between interpretability and accuracy, ultimately achieving robust 3D-aware spatial reasoning.

4 Experiments

4.1 Experimental Setup

Implementation Details. We use Qwen3-VL-8B-Instruct [3] as the base model for all experiments. Training is conducted in three stages: large-scale SFT (Stage 1), CoT cold-start training (Stage 2), and RL (Stage 3).

In Stage 1, the model is trained with a learning rate of 1×10^{-5} on a diverse mixture of datasets including the combined set of Levels I–III from 60% of the scenes, comprising 313.8K QA pairs, additional open-source spatial reasoning and GQA datasets [1, 11, 13, 19, 21, 28, 32, 48, 49]. In Stage 2, we perform cold-start training on 19.0K CoT annotated Level III samples with a learning rate of $1e-6$. In Stage 3, we employ GRPO algorithm for reinforcement learning on 24.8K high-level reasoning samples, using a learning rate of $1e-6$, with 8 rollouts per prompt.

For Stage 1 and Stage 2, the global batch size is set to 256, with a warmup ratio of 0.1. In Stage 3, a warmup ratio of 0.01 is used. All remaining training details, including dataset composition ratios and optimization hyperparameters, are provided in the Appendix.

Benchmarks. To evaluate the effectiveness of our proposed method in spatial reasoning, particularly under multi-view conditions, we conduct extensive experiments across several representative benchmarks. These include CV-Bench [35] for foundational single-view spatial perception, MMSI-Bench [45] and ViewSpatial-Bench [22] for multi-viewpoint spatial localization and reasoning, and 3DSR-Bench [27] for comprehensive 3D-aware spatial inference. This diverse selection of benchmarks allows for a holistic assessment of our model’s performance improvements across different levels of spatial complexity.

Baselines. We evaluate our method against a diverse range of MLLMs, including proprietary MLLMs (*e.g.*, GPT-5 [34]), representative open-source models across various parameter scales (*e.g.*, InternVL3-38B [50]), and specialized spatial reasoning MLLMs (*e.g.*, SPAR [48]). For baselines without publicly available scores on specific benchmarks, we evaluate their performance under the same experimental conditions as our own model to ensure a fair and rigorous comparison.

4.2 Main Results

Table 1: Performance of different models across four spatial reasoning benchmarks. The best overall results in each column are **bolded**, while the second-best overall results are *italicized*. The best results among open-source models are underlined.

Model	MMSI-Bench													CV	VS	3DSR
	Attribute		Motion (M)		Positional Relationship (PR)							MSR	Avg.			
	Appr.	Meas.	C	O	C-C	C-O	C-R	O-O	O-R	R-R						
<i>Proprietary MLLMs</i>																
GPT-5 [34]	36.40	60.90	32.40	<i>36.80</i>	<i>43.00</i>	48.80	51.80	35.10	42.40	32.10	42.40	41.90	-	-	-	66.70
GPT-4o [18]	<i>37.90</i>	37.50	37.80	31.60	37.60	20.90	33.70	25.50	32.90	27.20	32.80	32.10	78.90	34.98	-	60.30
Gemini2.5-Pro [9]	33.30	60.90	27.00	28.90	45.20	38.40	45.80	35.10	41.20	32.10	33.30	37.60	-	-	-	64.30
Seed1.5-VL [16]	21.20	35.90	37.80	27.60	27.60	20.90	28.90	30.90	40.00	27.20	30.80	30.70	-	-	-	64.00
<i>Open-source MLLMs</i>																
InternVL3-78B [50]	21.20	40.60	33.80	31.60	33.30	18.60	24.10	22.30	36.50	32.10	27.80	28.90	-	-	-	61.30
InternVL2.5-78B [7]	22.70	35.90	20.30	34.20	35.50	30.20	42.20	20.20	40.00	<i>38.30</i>	29.30	31.50	-	-	-	58.00
Qwen2-VL-72B-Instruct [37]	27.30	42.20	33.80	25.00	21.50	24.40	27.70	31.90	41.20	<i>38.30</i>	33.30	31.50	-	-	-	57.50
LLaVA-OneVision-72B [21]	18.20	32.80	16.20	28.90	40.90	30.20	30.10	25.50	35.30	32.10	16.70	26.90	-	-	-	57.90
InternVL3-38B [50]	30.30	42.20	20.30	32.90	25.80	23.30	28.90	24.50	37.60	29.60	28.30	29.00	-	-	-	59.10
Ovis2-34B [2]	31.80	39.10	20.30	32.90	31.20	26.70	21.70	27.70	<i>44.70</i>	32.10	31.30	30.80	-	-	-	58.60
Qwen2.5-VL-32B-Instruct [4]	36.40	39.10	27.00	34.20	26.90	26.70	33.70	26.60	30.60	29.60	21.70	28.90	-	-	-	56.90
InternVL3-14B [50]	24.20	34.40	27.00	27.60	15.10	27.90	22.90	29.80	35.30	24.70	26.80	26.70	-	40.28	-	56.00
InternVL3-8B [50]	22.70	32.80	17.60	31.60	21.50	30.20	37.30	27.70	43.50	30.90	30.80	29.90	85.50	-	-	55.50
InternVL3.5-8B [38]	19.69	35.93	17.56	26.31	21.50	18.60	28.91	24.46	29.41	19.75	28.28	24.9	82.86	43.24	-	-
MiMo-VL-7B-RL [41]	24.20	25.00	28.40	23.70	25.50	22.10	27.70	29.80	45.90	30.90	32.80	30.30	82.30	-	-	60.10
MiMo-VL-7B-SFT [41]	19.70	37.50	17.57	23.68	27.96	25.58	21.69	34.04	31.76	35.80	<i>34.34</i>	29.00	82.33	40.37	-	59.31
Qwen2.5-VL-7B-Instruct [4]	21.20	34.40	23.00	31.60	24.70	29.10	28.90	21.30	35.30	23.50	<u>25.30</u>	26.80	73.00	36.85	-	53.20
<i>Spatial-reasoning MLLMs</i>																
SPAR [48]	-	-	-	-	-	-	-	-	-	-	-	-	79.91	-	-	57.48
VST-7B-RL [44]	26.60	<i>48.50</i>	33.80	31.60	35.50	38.30	30.90	54.60	35.30	47.00	22.70	35.30	<i>86.50</i>	-	-	60.10
<i>Ours</i>																
Base Model (Qwen2-VL-72B-Instruct [3])	19.70	39.06	24.32	26.32	19.35	27.91	37.35	29.79	38.82	28.40	29.80	29.20	84.31	40.51	-	59.98
MV-STRIDE-SFT (Full)	43.94	43.75	45.95	31.58	34.41	66.28	56.63	<i>41.49</i>	34.12	30.86	22.73	<i>38.90</i>	86.73	48.35	-	64.51
MV-STRIDE-SFT (Stage 1)	27.27	29.69	<i>40.54</i>	34.21	<i>43.00</i>	<i>63.95</i>	66.27	36.17	34.12	28.40	21.72	37.20	86.39	50.28	<i>65.52</i>	-
MV-STRIDE-ColdStart (Stage 2)	30.30	31.25	<i>40.54</i>	35.53	40.86	56.98	<i>61.45</i>	39.36	29.41	28.40	23.74	36.70	85.94	47.30	-	61.08
MV-STRIDE-RL (Stage 3)	31.82	32.81	<i>40.54</i>	40.79	39.78	62.79	59.04	38.30	34.12	27.16	22.73	37.50	85.94	<i>49.30</i>	-	59.43

Quantitative Results on Spatial Reasoning Benchmarks. In this section, we evaluate the performance of our models across different training phases: MV-STRIDE-SFT(Full), a comprehensive baseline trained by pooling all data from all three stages for SFT; MV-STRIDE-SFT(Stage 1), the model obtained after the SFT stage; MV-STRIDE-ColdStart(Stage 2), the model obtained after the cold start stage; and MV-STRIDE-RL(Stage 3), the final version further refined through RL, comparing them against various MLLMs.

As illustrated in Tab. 1, our models demonstrate superior performance across multiple spatial reasoning benchmarks. Specifically, MV-STRIDE-SFT (Full) achieves state-of-the-art (SOTA) results among open-source MLLMs with an average accuracy of **38.9%** on the MMSI-Bench, yielding a substantial **9.7%** improvement over the Base Model. This performance surpasses several models with significantly larger parameter scales and remains highly competitive with leading proprietary MLLMs, effectively narrowing the gap between open-source models and closed-source counterparts in complex spatial reasoning.

A more granular analysis of MMSI-Bench sub-categories reveals that our models excel in dimensions highly correlated with our proposed dataset, such as Motion and Positional Relationship. Notably, it attains SOTA performance across all tested models in several challenging sub-tasks, including M-C and PR-C-O *et al.* These results provide compelling evidence for the effectiveness of our data in addressing complex multi-view spatial reasoning challenges.

Furthermore, our model exhibits remarkable generalization capabilities. MV-STRIDE-SFT(Stage 1) achieves the overall best performance on ViewSpatial-Bench and the top result among open-source models on 3DSRBench, both of which emphasize 3D and multi-view spatial reasoning. On the single-view CV-Bench, MV-STRIDE-SFT(Full) reaches the best performance among all compared models. Collectively, these empirical findings strongly validate that our proposed dataset significantly augments the model’s spatial reasoning capacity, particularly in synthesizing multi-view information for comprehensive 3D spatial understanding.

We further validate the effectiveness of MV-STRIDE on MLLMs with different parameter scales and architectures, and observe consistent improvements across benchmarks. Detailed results are provided in Appendix.

Qualitative Analysis of 3D-Consistent Reasoning. While the quantitative benchmarks of MV-STRIDE-RL may not always surpass our other variants, it exhibits a significant advantage in generating highly interpretable and structured spatial reasoning chains. As illustrated in Fig. 3, through cold-start fine-tuning and RL, the model has autonomously evolved a multi-stage systematic paradigm for addressing multi-view challenges.

This hierarchical process mirrors the human spatial cognition path: it begins with Single-View Perception, identifying localized assets like kitchen appliances; progresses to Cross-View 3D Modeling, establishing spatial anchors between frames and estimating camera pose (*e.g.*, quantifying horizontal shifts



Fig. 3: CoT produced by MV-STRIDE-RL on several multi-view spatial reasoning examples from MMSI-Bench.

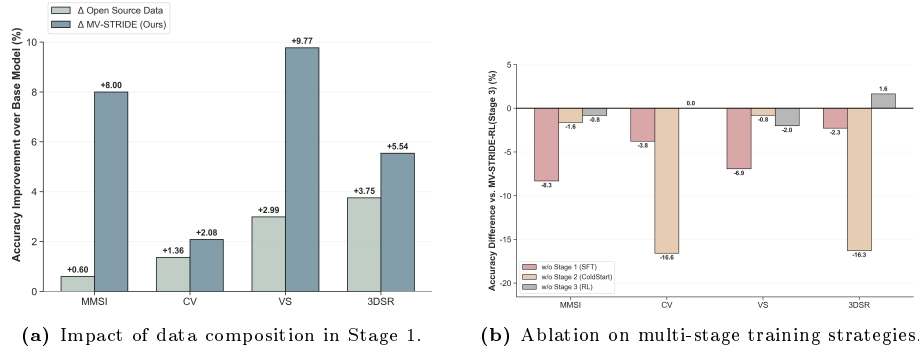


Fig. 4: Ablation study results. Results are reported in percentage points (pp.) relative to corresponding baselines.

and rotations); and culminates in High-Level Scene Reasoning to derive the final answer.

The samples in Fig. 3 demonstrate this qualitative leap: the model does not merely predict a label but constructs a consistent internal 3D representation, explicitly calculating the camera’s translation (*e.g.*, 0.85m shift) and rotation (*e.g.*, 30.5°). This level of mathematical grounding and logical coherence underscores the capacity of reinforcement learning to cultivate internal spatial consistency—providing a "white-box" reasoning process that exceeds the reach of standard supervised learning.

4.3 Ablation Studies

Effectiveness of MV-STRIDE. To investigate the specific contribution of our proposed dataset to the model’s spatial reasoning capabilities, we conduct a series of ablation experiments on the training data composition in Stage 1. As

established in our training pipeline, we introduce various open-source datasets including spatial reasoning, caption and GQA data as auxiliary data to fortify the model’s fundamental perception. However, to demonstrate that MV-STRIDE is the primary driver for the leap in spatial reasoning—especially in multi-view scenarios—we compare the performance of models trained with different data combinations.

As summarized in Fig. 4a, incorporating only open-source auxiliary data yields marginal improvements over the base model across most benchmarks, with almost negligible gains on the multi-view-focused MMSI-Bench. In stark contrast, the inclusion of MV-STRIDE leads to a substantial performance surge. Specifically, accuracy on MMSI-Bench and ViewSpatial experiences a significant leap, far outstripping the improvements brought by general open-source data. Furthermore, consistent and marked enhancements are observed across the other two benchmarks. These empirical results provide robust evidence that our proposed dataset is the essential catalyst for empowering the model to synthesize complex multi-view information and achieve superior 3D spatial understanding.

Effect of Progressive Multi-Stage Training. To verify the necessity of each phase in our proposed pipeline, we evaluate models at different training stages and perform an ablation study on every stage. The results are summarized in Fig. 4b and Tab. 1.

Necessity of Individual Stages. We conduct an ablation study by systematically removing stages, with Fig. 4b illustrating the performance delta relative to MV-STRIDE-RL (Stage 3). Without Stage 1’s foundational spatial knowledge, the model lacks a viable exploration space for reinforcement learning, leading to convergence failure on high-order tasks. Omitting the Stage 2 cold-start prevents the internalization of structured reasoning; during inference, these models suffer significant degradation on CV-Bench and 3DSRBench, proving that RL requires supervised warm-up to develop coherent logic. Crucially, without Stage 3, the cold-start model merely imitates SFT templates through rote memorization. In contrast, the RL phase empowers the model to autonomously verify spatial logic and perspective consistency, facilitating a transition from passive pattern matching to active, generalized problem-solving (see Appendix for qualitative comparisons).

Progressive Integration and the Reasoning Tax. Comparing the stages of progressive training in Tab. 1, we observe a slight decline in raw accuracy on certain benchmarks after transitioning from Stage 1 to Stage 2 and Stage 3.

During the transition from Stage 1 to Stage 2, the model changes from the short-answer format used in foundational SFT to CoT-style supervision. This shift alters the output distribution and may reduce direct-answer accuracy on some benchmarks, which we refer to as a "reasoning tax". From Stage 2 to Stage 3, the effect becomes benchmark- and task-dependent: multi-view-oriented benchmarks such as MMSI-Bench and ViewSpatial-Bench benefit more from RL training, whereas single-view or more general 3D reasoning benchmarks may remain stable or slightly decrease due to the focus of Stage 3 on multi-view

high-level reasoning. Despite these fluctuations, the multi-stage training process enables the model to internalize a more systematic reasoning pattern, improving interpretability and supporting cross-view spatial reasoning.

Multi-Stage Strategy vs. All-in-one SFT. We compare our multi-stage strategy with MV-STRIDE-SFT (Full), which pools all multi-stage data for a single-pass SFT in Tab. 1. While MV-STRIDE-SFT (Full) achieves higher accuracy on certain benchmarks such as MMSI-Bench, this is largely due to its stronger pattern-matching capability from seeing all CoT data as direct supervision. However, our multi-stage approach remains competitive and even superior in specific 3D-heavy contexts. More importantly, our approach ensures that the model’s reasoning is actively derived through RL optimization rather than just passively imitated. This leads to a model that is more robust to novel, out-of-distribution spatial scenarios where simple rote memorization of SFT data fails.

Table 2: Ablation study on data sources for Level III spatial reasoning. Performance is evaluated on MMSI-Bench (Acc.). "# Samples" indicates the volume of Level 3 data used for SFT.

Data Source	# Infinigen	# ScanNet++	MMSI-Bench
Base Model	-	-	29.20
Synthetic Only	99.3k	-	36.70
Real-world Only	-	99.5k	37.50
Hybrid	49.7k	49.8k	38.00
Unified	99.3k	99.5k	39.40

Impact of Synthetic vs. Real-world Data Sources. This section investigates the complementarity of synthetic and real-world data sources for training spatial reasoning models. As illustrated in Tab. 2, we conduct supervised fine-tuning using equal quantities of Level 3 spatial reasoning data from different sources to evaluate their respective impacts.

The experimental results reveal that although synthetic data sources like Infinigen [31] offer exceptional generation flexibility and theoretical reproducibility, their diversity remains limited by the underlying procedural generation algorithms. Additionally, the fidelity gap compared to real-world scenes causes the performance gains from synthetic data to be slightly lower than those from real-world sources such as ScanNet++ [46]. Nevertheless, the hybrid use of both synthetic and real-world data yields the optimal performance. This synergy stems from the fact that synthetic data provides engine-generated ground truth with absolute precision, effectively helping the model establish rigorous spatial geometric concepts. Concurrently, real-world data accounts for the complexity of visual distributions, ensuring that the acquired spatial reasoning capabilities generalize effectively to real-world inputs. These findings demonstrate the ne-

cessity of combining controllable simulation with high-fidelity real-world data for Development of large-scale spatial reasoning datasets.

5 Conclusion

This paper presents MV-STRIDE, a hierarchical dataset and training framework designed to bridge the gap in multi-view spatial reasoning for MLLMs. By explicitly modeling capability dependencies—from foundational perception to complex 3D reasoning—we provide a cognitively grounded learning pathway that significantly enhances spatial inference compared to conventional "flat" datasets. Our evaluations demonstrate that while supervised fine-tuning establishes a robust performance baseline, the incorporation of reinforcement learning is essential for cultivating internal spatial consistency. Notably, we find that RL-refined models autonomously develop structured, interpretable reasoning chains, moving beyond rote memorization toward genuine 3D spatial logic. Furthermore, our results underscore that the synergy between precise synthetic grounding and real-world visual complexity is critical for balancing geometric accuracy with robustness. Ultimately, MV-STRIDE offers a systematic roadmap for advancing MLLMs from isolated 2D perception toward holistic, multi-view consistent spatial cognition.

References

1. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8076–8084 (2019)
2. AIDC-AI: Ovis2. Open-source Multimodal Large Language Model (2025), <https://github.com/AIDC-AI/Ovis>, accessed: 2026-03-02
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 9490–9498. IEEE (2025)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)

8. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. In: *Advances in Neural Information Processing Systems*. vol. 37, pp. 135062–135093 (2024)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025)
10. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., Grasch, P.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7395–7408 (2025)
11. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muenighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 91–104 (2025)
12. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Salvador, J., Ehsani, K., Han, W., Kolve, E., Farhadi, A., Kembhavi, A., Mottaghi, R.: Proctor: Large-scale embodied ai using procedural generation. In: *Advances in Neural Information Processing Systems*. vol. 35, pp. 5982–5994 (2022)
13. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., Xu, H., Theiss, J., Chen, T., Li, J., Tu, Z., Wang, Z., Ranjan, R.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2026)
14. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: *European Conference on Computer Vision*. pp. 148–166. Springer (2024)
15. Google: Gemini 3. Online AI assistant (2025), <https://gemini.google.com/>, accessed: 2026-03-02
16. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062* (2025)
17. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006* (2025)
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
19. Jung, J.H., Kim, E.T., Kim, S., Lee, J.H., Kim, B., Chang, B.: Is ‘right’ right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14257–14267 (2025)
20. Lee, P.Y., Je, J., Park, C., Uy, M.A., Guibas, L., Sung, M.: Perspective-aware reasoning in vision-language models via mental imagery simulation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9241–9251 (2025)
21. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research* (2024)

22. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. arXiv preprint arXiv:2505.21500 (2025)
23. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: SpatialLadder: Progressive Training for Spatial Reasoning in Vision-Language Models. In: International Conference on Learning Representations (2026)
24. Liu, Y., Ma, M., Yu, X., Ding, P., Zhao, H., Sun, M., Huang, S., Wang, D.: Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In: Advances in Neural Information Processing Systems. vol. 38, pp. 123926–123958 (2025)
25. Liu, Y., Chi, D., Wu, S., Zhang, Z., Hu, Y., Zhang, L., Zhang, Y., Wu, S., Cao, T., Huang, G., et al.: Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. arXiv preprint arXiv:2501.10074 (2025)
26. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al.: Ovis2. 5 technical report. arXiv preprint arXiv:2508.11737 (2025)
27. Ma, W., Chen, H., Zhang, G., Chou, Y.C., de Melo, C.M., Yuille, A., Chen, J.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)
28. Ma, W., Chou, Y.C., Liu, Q., Wang, X., de Melo, C., Xie, J., Yuille, A.: Spatial-reasoner: Towards explicit and generalizable 3d spatial reasoning. In: Advances in Neural Information Processing Systems. vol. 38, pp. 140751–140774 (2025)
29. OpenAI: GPT-4o. Online AI assistant (2024), <https://openai.com/index/gpt-4o-system-card/>, accessed: 2026-03-02
30. OpenAI: GPT-5. Online AI assistant (2025), <https://openai.com/index/introducing-gpt-5/>, accessed: 2026-03-02
31. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12630–12641 (2023)
32. Ray, A., Duan, J., Brown, E., Tan, R., Bashkurova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., Saenko, K.: Sat: Dynamic spatial aptitude training for multimodal language models (2025)
33. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
34. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
35. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
36. Wang, J., Zhang, Z., Liu, Z., Li, Y., Ge, J., Xie, H., Zhang, Y.: Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
37. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

38. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
39. Wang, X., Ma, W., Zhang, T., de Melo, C.M., Chen, J., Yuille, A.: 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. pp. 24669–24679 (2025)
40. Wang, Y., Ji, Y., Liu, Y., Zhou, E., Yang, Z., Tian, Y., Qin, Z., Liu, Y., Tan, H., Chi, C., et al.: Towards cross-view point correspondence in vision-language models. arXiv preprint arXiv:2512.04686 (2025)
41. Xiaomi, L.C.T.: Mimo-vl technical report (2025)
42. Xu, R., Wang, W., Tang, H., Chen, X., Wang, X., Chu, F.J., Lin, D., Feiszli, M., Liang, K.J.: Multi-spatialmlm: Multi-frame spatial understanding with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
43. Xue, Q., Liu, W., Wang, S., Wang, H., Wu, Y., Gao, W.: Reasoning path and latent state analysis for multi-view visual spatial reasoning: A cognitive science perspective. arXiv preprint arXiv:2512.02340 (2025)
44. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
45. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. International Conference on Learning Representations (2026)
46. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
47. Zhan, X., Huang, W., Sun, H., Fu, X., Ma, C., Cao, S., Jia, B., Lin, S., Yin, Z., Bai, L., Ouyang, W., Li, Y., Guo, J., Guo, Y.: Actial: Activate spatial reasoning ability of multimodal large language models. In: Advances in Neural Information Processing Systems. vol. 38, pp. 83946–83971 (2025)
48. Zhang, J., Chen, Y., Xu, Y., Huang, Z., Mei, J., Chen, C., Zhou, Y., Yuan, Y.J., Cai, X., Huang, G., Quan, X., Xu, H., Zhang, L.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. In: Advances in Neural Information Processing Systems. vol. 38 (2025)
49. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., Zhang, S.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In: Advances in Neural Information Processing Systems. vol. 38, pp. 28404–28481 (2025)
50. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)