ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models

Dingming Li[*1], Hongxing Li[*1], Zixuan Wang[1], Yuchen Yan[1], Hang Zhang[1], Siqi Chen[1], Guiyang Hou[1], Shengpei Jiang[2], Wenqi Zhang[1], Yongliang Shen[†1], Weiming Lu[1], and Yueting Zhuang[1]

[1] Zhejiang University
[2] The Chinese University of Hong Kong
dingmingli78@gmail.com, syl@zju.edu.cn

Abstract. Vision-language models (VLMs) have demonstrated remarkable capabilities in understanding and reasoning about visual content, but significant challenges persist in tasks requiring cross-viewpoint understanding and spatial reasoning. We identify a critical limitation: current VLMs excel primarily at egocentric spatial reasoning (from the camera's perspective) but fail to generalize to allocentric viewpoints when required to adopt another entity's spatial frame of reference. We introduce **ViewSpatial-Bench**, a dedicated multi-perspective spatial localization benchmark covering both camera-perspective and human perspective reasoning, supported by an automated 3D annotation pipeline that generates precise directional labels. Comprehensive evaluation of diverse VLMs on ViewSpatial-Bench reveals a significant performance disparity: models demonstrate reasonable performance on camera-perspective tasks but exhibit reduced accuracy when reasoning from a human viewpoint. By fine-tuning VLMs on our multi-perspective spatial dataset, we achieve an overall performance improvement of 46.24% across tasks, confirming the strong latent spatial reasoning capacity of VLMs under targeted supervision. Our work establishes a crucial benchmark for spatial intelligence in embodied AI systems and provides empirical evidence that modeling 3D spatial relationships enhances VLMs' corresponding spatial comprehension capabilities.Our benchmark is available at `HuggingFace`.

Keywords: Multi-viewpoint Spatial Reasoning · Vision-Language Models · Spatial Benchmark Evaluation

1 Introduction

While Vision-Language Models (VLMs) demonstrate remarkable capabilities in visual content understanding and reasoning [3, 6, 30], they exhibit significant limitations when confronted with complex tasks requiring cross-viewpoint comprehension and spatial reasoning [29, 31]. Specifically, current VLMs perform adequately in egocentric spatial judgments but struggle to interpret and reason

[*] Equal contribution, [†] Corresponding author.

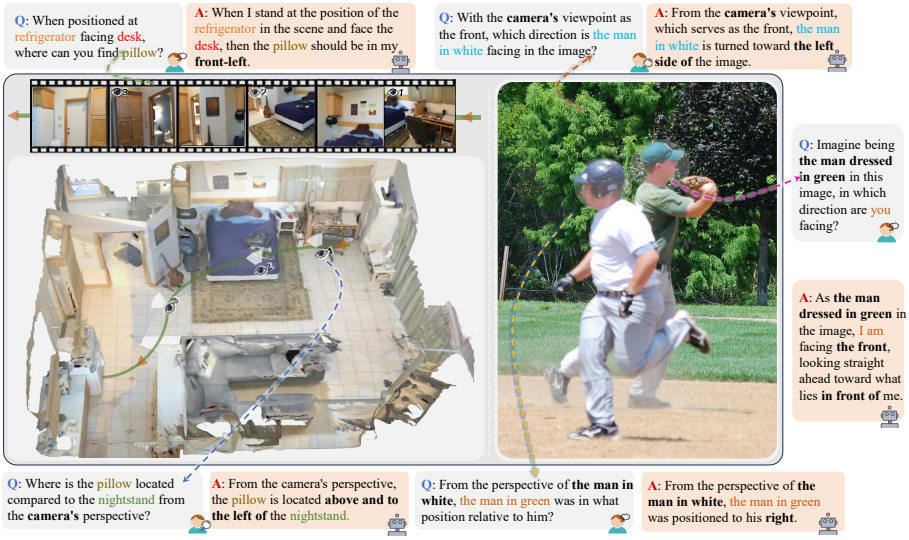

Fig. 1: ViewSpatial-Bench for multi-perspective spatial reasoning. Our benchmark evaluates spatial localization capabilities from both camera and human perspectives across five task types.

about spatial relationships from alternative entity perspectives [17]. This constraint substantially impedes the performance of the model in practical application scenarios.

Humans naturally understand spatial relationships from multiple perspectives. When interacting with others, we effortlessly adopt their viewpoints to interpret spatial references: intuitively distinguishing between "*the cup on my left*" and "*the cup on your left*" without conscious effort. This perspective-taking ability enables seamless communication in physical spaces and forms the foundation for successful collaborative interactions. In contrast, current VLMs operate primarily within an egocentric reference frame, where spatial reasoning is entirely anchored to the camera's perspective [26].

This issue is particularly prominent in embodied interaction scenarios. When a person asks a robot "Can you pass the mug on my right?", they expect the robot to identify the target object from their perspective rather than the robot's own. This ability to reason spatially from different viewpoints, known in cognitive science as "perspective-taking," represents a critical capability for human-machine interaction, spatial navigation [47], and multi-agents collaboration [12]. Crucially, this challenge becomes significantly more complex in three-dimensional environments, where viewpoint transformation involves not only changes in two-dimensional planes but also considerations of depth, occlusion, and camera pose, factors that substantially increase the difficulty of object localization tasks [20].

Currently, most VLMs rely primarily on large-scale image-text pairs harvested from the webs, where spatial information tends to be sparse due to the inherent

lack of three-dimensional spatial annotations [23]. Moreover, even in multimodal datasets that include spatial descriptions, task designs typically remain limited to shallow spatial understanding from static viewpoints, lacking multi-dimensional, multi-perspective spatial reasoning tasks that would enable models to develop more generalizable spatial representations [6, 44]. We therefore hypothesize that VLMs' deficiencies in cross-viewpoint spatial understanding tasks stem from structural limitations in their training data.

To address this research gap, we introduce ViewSpatial-Bench, a dedicated benchmark for multi-perspective spatial localization, covering both camera-perspective and human (allocentric) perspective reasoning. This benchmark encompasses five distinct localization recognition tasks and is supported by a reliable automated 3D orientation annotation pipeline that generates efficient, diverse, and scalable image datasets with precise directional labels.Furthermore, we utilized this automated pipeline to produce extensive spatially annotated training data for VLMs, enhancing their perceptual reasoning capabilities for spatial relationships across multiple viewpoints.

Based on ViewSpatial-Bench, we conducted a comprehensive evaluation of multiple VLMs investigating their spatial understanding performance. Results demonstrate significant limitations in spatial localization tasks, particularly when reasoning across different viewpoints. To address these limitations, we introduced well-annotated spatial data for VLM training, enabling more concrete multi-perspective spatial understanding and yielding the Multi-View Spatial Model. This approach significantly improved spatial perception across viewpoints, partially validating our hypothesis. In summary, our contributions are:

- We propose **ViewSpatial-Bench**, a dedicated benchmark for evaluating multi-viewpoint spatial localization across 5,700 curated samples and five task types. This benchmark systematically assesses VLMs' spatial reasoning from both camera and human perspectives, addressing a critical gap in cross-viewpoint evaluation frameworks;
- We design an automated 3D spatial annotation pipeline that efficiently generates large-scale, precisely annotated multi-view datasets. This pipeline provides rich spatial relationship data for VLM training through automated orientation annotation, establishing important foundations for future research;
- We develop the Multi-View Spatial Model trained on our large-scale multi-viewpoint VQA dataset, and identify fundamental limitations in current VLMs' perspective-based spatial reasoning. Fine-tuning on our dataset yields a 46.24% improvement over baselines, confirming the strong latent spatial reasoning capacity of VLMs under targeted supervision.

2 Related Works

Spatial Reasoning with VLMs. Recently, VLMs have demonstrated significant advancements in understanding and reasoning about visual content [2, 10]. Both proprietary and open-source models have achieved impressive performance in

visual question answering, image captioning, and complex multimodal reasoning tasks. These models typically incorporate image encoders and vision-language fusion modules [7, 19, 22], pre-trained on large-scale image-text pairs [43].

However, despite current VLMs' exceptional performance on certain visual reasoning tasks, their spatial understanding capabilities remain fundamentally limited [6, 29]. When handling tasks involving spatial relationships, object localization, or embodied interaction reasoning, models typically rely on camera-centric reference frames, with their spatial understanding strictly bound to the observational viewpoint [29, 39]. This constraint impairs their generalization capabilities and practical utility in tasks requiring perspective transformation or third-person spatial comprehension, making the development of models with stronger perspective-taking awareness a critical challenge for advancing multimodal intelligence.

Benchmarks fo Spatial Perspective-Taking. Several benchmarks have been proposed to evaluate spatial reasoning capabilities in VLMs, but most focus primarily on single-perspective spatial understanding. For instance, EmbSpatial-Bench [11] and What'sUP [16] concentrate on assessing models' abilities to recognize spatial relationships between objects in two-dimensional images, while VSI-Bench [39] tests model performance on compositional visual reasoning tasks involving spatial queries. Additionally, some research explores spatial reasoning in embodied AI, such as navigation and object localization tasks, but these works predominantly rely on the agent's egocentric perspective [30].

Although some benchmarks have begun to address cross-viewpoint spatial understanding, such as 3DSRBench [24] and SPHERE [45], they remain insufficient in terms of multi-task comprehensiveness and depth of perspective transformation assessment.

3 ViewSpatial-Bench

3.1 Overview

We introduce ViewSpatial-Bench to quantitatively evaluate VLMs' spatial localization capabilities in 3D environments from multiple perspectives. Our benchmark contains over 5,700 question-answer pairs spanning more than 1,000 unique 3D scenes, with source imagery from the validation sets of ScanNet [9] and MS-CoCo [21]. Following a construction pipeline illustrated in Figure 2, we first acquired images with complete spatial information, created metadata using existing annotations, extracted spatial relationships for specific tasks, and finally constructed and filtered the QA dataset.

ViewSpatial-Bench comprises five localization recognition tasks across two complementary perspective frameworks. From the camera perspective: (1) Object Relative Direction recognition(Cam-Rel. Dir.), which determines spatial relationships between objects directly from images; (2) Object View Orientation recognition(Cam-Obj. Ori.), which identifies the gaze direction of individuals relative to the camera from an egocentric viewpoint. These tasks evaluate VLMs'

intuitive, egocentric spatial understanding abilities. From the human perspective: (3) Object Relative Direction recognition(Per-Rel. Dir.), which involves adopting the viewpoint of a character in the image to determine the spatial relationships of other objects from their perspective; (4) Object View Orientation recognition(Per-Obj. Ori.), which requires assuming the position of a character in the image to determine the direction of their gaze; (5) Scene Simulation Relative Direction recognition(Per-Sce. Sim.), which requires modeling oneself within a spatial scene across sequential frames to determine relative positions of other objects. These latter three tasks assess VLMs' abstract, perception-dependent spatial awareness while accommodating complex human pose variations and spatial information in embodied scenarios.

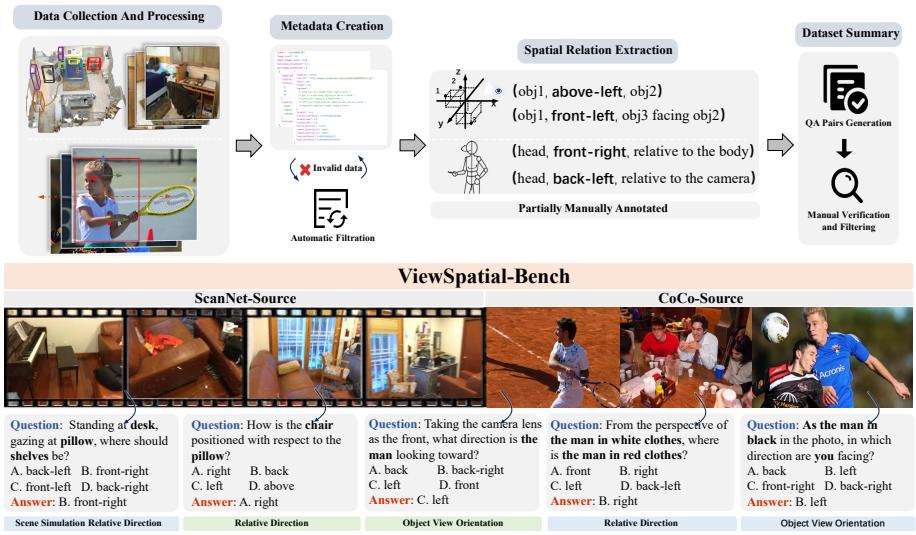

Fig. 2: ViewSpatial-Bench construction pipeline. From data collection to QA generation across camera perspective ( ) and person perspective ( ) tasks. The pipeline includes metadata creation, automatic filtering, spatial relation extraction, and manual verification.

3.2 Dataset Construction

ViewSpatial-Bench construction follows a systematic process using two complementary data sources: ScanNet for rich 3D scene reconstructions with accurate spatial coordinates, and MS-CoCo for diverse images with human subjects and annotated keypoints. This combination supports both precise 3D spatial reasoning and perspective-dependent person-centric understanding tasks. We developed specialized processing pipelines for each source to extract reliable spatial relationships using automated techniques with manual verification.

ScanNet Source. For Cam-Rel. Dir. and Per-Sce. Sim. tasks, we utilized the ScanNet validation set. We first obtained voxel information for each scene, then applied Maximum Coverage Sampling (Algorithm 1 [48]) to ensure complete spatial representations with minimal frames while maximizing diversity. This approach prevented redundant capture of the same spatial locations. For each selected frame, we generated scene metadata including visible objects with visibility rates and 3D spatial coordinates in the camera coordinate system.

For Cam-Rel. Dir. task, we leveraged 3D spatial coordinates and camera parameters to determine relative positions between object pairs. For Per-Sce. Sim. task, we first identified objects appearing only once in each scene (set $\mathcal{N}$), selected object triads o_1, o_2, o_3 from $\mathcal{N}$, and used metadata to locate frames containing all three objects. By simulating the position and orientation at o_1, we calculated the relative position of o_3 from this simulated viewpoint.

MS-CoCo Source. For Cam-Obj. Ori. and Per-Obj. Ori. tasks, plus Per-Rel. Dir. task, we utilized the MS-CoCo validation set. We filtered images containing animate objects occupying at least 20% of the image area.

Algorithm 1 Maximum Coverage Sampling

Require: Set of frames $F = \{f_1, f_2, \ldots, f_n\}$, voxel sets V_k for each frame f_k, budget K
Ensure: Subset $S \subseteq F$ maximizing voxel coverage

Initialize $S \leftarrow \emptyset$
Initialize $U \leftarrow \emptyset$ {Covered voxels set}
while size of S is less than K **do**
  Select $f^* = \arg\max_{f_k \in F \setminus S} |V_k \setminus U|$
  Add f^* to S
  Update $U \leftarrow U \cup V_{f^*}$
  if Stop condition is met **then**
    break
  end if
end while
return S

Algorithm 2 Head-to-body Orientation Offset

Require: Image I, keypoints K, bounding box B, Orient-Anything model D
Ensure: Person gaze direction

$P \leftarrow \text{Crop}(I, B)$
$(L_x, L_y), (R_x, R_y) \leftarrow \text{ExtractShoulders}(K)$
if $\text{Visibility}(L_y) = 0$ OR $\text{Visibility}(R_y) = 0$ **then**
  return False
end if
$H \leftarrow \min(L_y, R_y)$
$P_{head} \leftarrow P[0 : H, :]$, $P_{body} \leftarrow P[H :, :]$
$(az_{head}, conf_{head}) \leftarrow D(P_{head})$
$(az_{body}, conf_{body}) \leftarrow D(P_{body})$
$\Delta \leftarrow (az_{head} - az_{body} + 540) \bmod 360 - 180$
return direction based on Δ thresholds for left, front-left, front, front-right, right

For orientation tasks, we selected images where subjects' gaze directions aligned with head orientations. Using MS-CoCo's bounding boxes and keypoints, we segmented person images into head and body components, then employed Orient-Anything-Large [36] to calculate rotation angles (Algorithm 2). For person-perspective orientation, we derived gaze direction by analyzing angular offsets between head and body orientations. For camera-perspective orientation, we calculated both head and body rotation angles, selecting the computation with

highest confidence. For complex cases with multiple subjects, we resorted to manual annotation.

For Per-Rel. Dir. task, which include questions like "From person A's perspective, where is person B located?", we manually annotated 864 instances due to the complexity of human and object appearances and insufficient accuracy in automated approaches.

QA Dataset Creation. ViewSpatial-Bench is structured as a multiple-choice benchmark derived systematically from our metadata. After extracting 3D spatial information through our ScanNet and MS-COCO processing pipelines, we converted the raw spatial coordinates and orientation angles into standardized directional relationships using a rule-based mapping system. For each task category, we designed question templates that explicitly test perspective transformation abilities. The construction followed three key steps:

First, we converted raw spatial data (3D coordinates, orientation angles) into standardized directional relationships using angle-based mapping (e.g., 22.5° to 67.5° as "front-right," 67.5° to 112.5° as "right"). This discretization enabled consistent labeling across different scenes.

Second, we populated templates with object identifiers and computed spatial relationships from our metadata. For complex spatial reasoning tasks, our templates incorporate three objects to test perspective adoption with relative positioning:

QA Generation Example

Template: "If you stand at object1 facing object2, where is object3?"

Metadata: bookshelf(1.2, 0.5, 0), window(1.2,3.5,0), sofa(3.2,1.5,0)

Computation:
1. Vector bookshelf→window: (0,3.0,0) [front direction]
2. Vector bookshelf→sofa: (2.0,1.0,0)
3. Angle: 63.43° clockwise = "front-right"

Question: "If you stand at the bookshelf facing the window, where is the sofa?"
Answer: "front-right"
Distractors: "left", "back", "front-left"

Finally, we implemented specific rules for distractor generation: for single-directional attributes (e.g., "front"), distractors exclude compound directions containing that attribute ("front-left"); for compound directions (e.g., "front-left"), distractors exclude constituent single directions ("front" or "left"). This design systematically eliminates ambiguity and provides focused assessment of fundamental spatial concepts while controlling for question difficulty.

Filtering and Human Verification. To ensure the quality of ViewSpatial-Bench, we implemented a multi-stage filtering process for all tasks. During

metadata generation, we eliminated invalid data with incorrectly calculated orientation angles or excessively large rotation angles. In the manual filtering stage, for relative direction tasks, we removed instances where objects were too close to each other, objects were difficult to identify, or images were blurry. For gaze direction recognition tasks, we filtered out data where subjects' gaze directions significantly differed from their head orientations or where subjects were difficult to identify. Following automated construction and filtering, we conducted manual verification to confirm that target objects were clearly visible in images and that the spatial localizations were correct and unambiguous. This iterative refinement process continued until ViewSpatial-Bench met our quality standards [11, 39]. Comprehensive details regarding dataset construction procedures and human annotation protocols are provided in Appendix B.1.

3.3 Dataset Statistics

Figure 3 illustrates the five task categories in ViewSpatial-Bench and their respective proportions. To ensure balanced evaluation across observational frameworks, camera perspective and person perspective tasks account for 48.4% and 51.6% of the benchmark respectively, enabling fair and direct comparison of spatial reasoning capabilities under each viewpoint. For the Relative Direction task from camera perspectives, which most directly probes 3D scene understanding, additional samples were incorporated to enrich the diversity of spatial configurations.

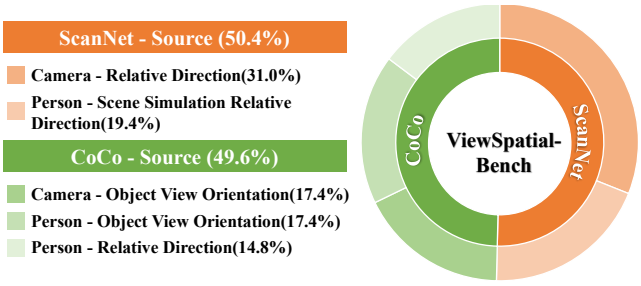

Fig. 3: Distribution of task categories in ViewSpatial-Bench, balanced between ScanNet-Source and CoCo-Source approaches, with five distinct subtasks for comprehensive evaluation of spatial reasoning across different viewpoints.

Table 1 presents a comprehensive comparison between ViewSpatial-Bench and existing spatial reasoning benchmarks. ViewSpatial-Bench contains 5,712 samples across 1,338 unique scenes, employing a hybrid construction method that combines automated 3D annotation pipelines with manual verification. The benchmark distinguishes itself with 18 distinct directional categories and precise 3D coordinate annotations from ScanNet, while uniquely supporting evaluation across both camera-perspective and person-perspective viewpoints for comprehensive assessment of perspective-taking capabilities for embodied AI applications. Detailed statistical analysis is provided in Appendix B.2.

Table 1: Comparison of ViewSpatial-Bench with existing spatial reasoning benchmarks. ViewSpatial-Bench is a dedicated benchmark for multi-perspective spatial localization, uniquely supporting both camera and person viewpoints with the broadest directional categories (18 directions in union across five task types) and spatial query targets.

Benchmark	Construct Method	Visual Diversity	Scale & Diversity		3D Annotation		Multi-Perspective		Spatial Query Target	
			Samples	Scenes	Directions	3D-Coord	Camera	Person	Person-Target	Object-Target
SpatialRGPT-Bench [6]	Automated	Single	1,410	524	6	✓	✓	✗	✗	✓
EmbSpatial-Bench [11]	Automated	Single	3,640	284	6	✓	✓	✗	✗	✓
What'sUP [16]	Manual	Single	820	205	12	✗	✓	✗	✗	✓
VSI-Bench [39]	Automated	Multi	3,672	245	8	✓	✓	✓	✗	✓
3DSRBench [24]	Manual	Single	2,772	1,827	8	✗	✓	✗	✓	✓
SPHERE [45]	Manual	Single	2,285	1,001	7	✗	✓	✓	✓	✓
All-Angles Bench [41]	Hybrid	Single	2,132	90	4	✗	✓	✗	✓	✓
MindCube [42]	Hybrid	Multi	1050	975	-	✗	✓	✗	✗	✓
GSR-BENCH [27]	Automated	Multi	820	205	12	✗	✓	✗	✗	✓
ViewSpatial-Bench	**Hybrid**	**Multi**	**5,712**	**1,338**	**18**	✓	✓	✓	✓	✓

4 Multi-View Spatial Fine-Tuning

To address the limitations in perspective-dependent spatial reasoning identified in current VLMs, we fine-tune existing models on our multi-perspective spatial dataset. For convenience, we refer to the resulting fine-tuned models as Multi-View Spatial Models (MVSM).

Following the ViewSpatial-Bench construction pipeline, we leveraged our automated spatial annotation framework to generate approximately 43K diverse spatial relationship samples across all five task categories. This dataset incorporates 3D spatial information from ScanNet [9] and MS-COCO [21] training sets, supplemented with Spatial-MM [29] data for the Person-perspective Relative Direction task where full automation proved challenging due to complex human spatial coordinates and environmental contexts.We structured all training data as image-text pairs using consistent natural language templates to articulate spatial relationships between objects or entities, with answers represented as standardized directional classifications. Our fine-tuning strategy trains models on mixed batches containing samples from both camera and human perspectives, encouraging the development of unified representations of 3D spatial relationships that support robust reasoning across different viewpoints.

5 Experiments

5.1 Experimental Setup

Baselines and Metrics. We conducted comprehensive evaluations of current VLMs on ViewSpatial-Bench using accuracy as our primary metric. Our evaluation includes a diverse set of models spanning different architectures and parameter scales: (1) Open-source models: InternVL2.5/VL3 [5,49], LLaVA-NeXT-Video [46], LLaVA-OneVision [18], Llama-3.2-Vision [13], Kimi-VL-Instruct [34],

Qwen2.5/3/3.5-VL [1, 35, 38]; (2) Proprietary models: GPT-4o [15], GPT-5-mini [25], Gemini-2.0-Flash [33], Gemini-2.5-Flash [8], Gemini-2.5-Pro [8], Gemini-3.0-Flash [32], GLM-4.6v [14], Doubao-Seed-1.8 [28], Doubao-Seed-2.0 [28];

Implementation Details. We employ our MVSM fine-tuning strategy on Qwen2.5-VL-3B [1] as the backbone model. Detailed training configurations and evaluation procedures are provided in Appendix C.1 and C.2.

Table 2: Zero-shot performance on ViewSpatial-Bench. Accuracy comparison across multiple VLMs on camera and human perspective spatial tasks. **Bold** and underlined entries denote the best and second-best results within each model family. Multi-View Spatial Fine-Tuning yields consistent performance gains across all task categories, demonstrating its effectiveness as a generalizable approach to spatial reasoning.

Model	Camera-based Tasks			Person-based Tasks				Overall
	Rel. Dir.	Obj. Ori.	Avg.	Obj. Ori.	Rel. Dir.	Sce. Sim.	Avg.	
Proprietary Models								
GPT-4o [15]	41.46	19.58	33.57	42.97	40.86	26.79	36.29	34.98
Gemini-2.0-Flash [33]	45.29	12.95	33.66	41.16	32.78	21.90	31.53	32.56
GPT-5-mini [25]	56.97	27.41	46.34	43.98	49.29	26.06	38.77	42.44
Gemini-2.5-Flash [8]	52.62	23.09	42.00	42.97	42.16	20.27	34.22	37.99
Gemini-2.5-Pro [8]	58.71	32.73	49.37	48.59	45.84	25.79	39.24	44.15
Gemini-3.0-Flash [32]	62.94	35.54	53.08	44.88	60.69	26.24	42.40	47.58
GLM-4.6v [14]	56.35	36.35	49.16	**48.90**	47.39	23.44	38.91	43.87
Doubao-Seed-1.8 [28]	62.10	**45.28**	56.05	44.98	62.47	**33.67**	45.74	50.74
Doubao-Seed-2.0 [28]	**65.60**	44.78	**58.11**	47.19	**72.09**	33.57	**49.20**	**53.52**
Open-Source General Models								
InternVL2.5 (2B) [5]	38.52	22.59	32.79	47.09	40.02	25.70	37.04	34.98
Qwen3-VL (4B) [38]	46.98	28.01	40.16	45.68	29.22	17.74	30.48	35.17
Qwen2.5-VL (7B) [1]	46.64	29.72	40.56	37.05	35.04	28.78	33.37	36.85
LLaVA-NeXT-Video (7B) [46]	26.34	19.28	23.80	44.68	38.60	29.05	37.07	30.64
LLaVA-OneVision (7B) [18]	29.84	26.10	28.49	22.39	31.00	26.88	26.54	27.49
InternVL2.5 (8B) [5]	49.41	**41.27**	46.48	46.79	42.04	32.85	40.20	43.24
Qwen3-VL (8B) [38]	54.60	30.32	45.87	45.28	35.75	26.79	35.61	40.58
Llama-3.2-Vision (11B) [13]	25.27	20.98	23.73	51.20	32.19	18.82	33.61	28.82
InternVL3 (14B) [49]	54.65	33.63	47.09	33.43	37.05	31.86	33.88	40.28
Kimi-VL-Instruct (16B) [34]	26.85	22.09	25.14	**63.05**	43.94	20.27	41.52	33.58
Qwen2.5-VL (32B) [1]	39.03	29.92	35.75	36.45	34.68	21.09	30.18	32.88
Qwen2.5-VL (72B) [1]	50.65	26.71	42.04	42.17	42.76	24.80	35.82	38.83
Qwen3-VL-Thinking (235B) [38]	59.73	36.95	51.54	43.67	48.93	31.67	40.67	45.94
Qwen3.5-Plus (397B) [35]	**62.21**	38.65	**53.74**	50.20	**68.17**	**38.37**	**50.90**	**52.28**
Multi-View Spatial Fine-Tuning								
Qwen2.5-VL (3B) [1]	43.43	33.33	39.80	39.16	28.62	28.51	32.14	35.85
+SFT	**83.59**	**87.65**	**85.05**	**90.16**	**71.14**	**75.75**	**79.31**	**82.09**
Improvement over backbone	+40.16	+54.32	+45.25	+51.00	+42.52	+47.24	+47.17	+46.24
Random Baseline	25.16	26.10	25.50	24.60	31.12	26.33	27.12	26.33

5.2 Main Results

As shown in Table 2, our comprehensive evaluation reveals critical insights into the spatial reasoning capabilities of current VLMs and validates our approach:

Fundamental limitations in perspective-based spatial reasoning: Even powerful proprietary models like GPT-4o (34.98%), Gemini-2.0-Flash (32.56%), and Gemini-2.5-Flash (37.99%) demonstrate surprisingly weak spatial localization capabilities, barely outperforming random chance (26.33%). This confirms our hypothesis presented in the introduction that current VLMs, despite their impressive performance on standard vision-language tasks, fundamentally struggle with perspective-dependent spatial reasoning. The consistently poor performance across diverse architectures suggests it is not merely an implementation issue but a systematic deficiency in how these models conceptualize spatial relationships.

Egocentric vs. allocentric reasoning gap: Most VLMs demonstrate consistently higher spatial localization accuracy from camera perspectives (averaging 41.6%) than from human viewpoints (averaging 37.3%). This aligns with findings in prior work, as egocentric spatial reasoning is generally considered more tractable than allocentric reasoning. A plausible explanation lies in the composition of web-harvested training data, where camera-perspective spatial compositions are substantially more prevalent, leading VLMs to implicitly encode spatial priors that favor egocentric viewpoints. Nevertheless, the persistent gap between the two perspectives highlights a fundamental limitation in current VLMs' perspective-taking ability, which humans naturally possess.

Task-specific performance asymmetries: A particularly revealing pattern emerges in the interaction between task type and perspective. Most VLMs perform significantly worse on Object View Orientation tasks from camera perspectives compared to Relative Direction tasks(30.0% for Object View Orientation vs. 48.2% for Relative Direction), yet under person perspectives, performance on both tasks converges, with Object View Orientation marginally surpassing Relative Direction. This striking asymmetry confirms our hypothesis that current VLMs lack consistent cross-viewpoint spatial understanding. The discrepancy suggests these models fail to construct a coherent 3D representation that can be flexibly navigated from different viewpoints, instead treating different perspective-task combinations as essentially separate problems.

Effectiveness of perspective-aware training: Our Multi-View Spatial Model achieves dramatic improvement compared to its backbone Qwen2.5-VL (3B) model, representing a 46.24% absolute performance gain. The model shows remarkably consistent improvements across all task categories. The most substantial gains occur in orientation tasks, with improvements of 54.32% for camera-perspective and 51.00% for person-perspective Object View Orientation tasks. This symmetrical improvement pattern is particularly noteworthy, as it demonstrates that explicit training on diverse spatial annotations with perspective awareness enables the development of unified 3D spatial representations that function effectively across viewpoints.

5.3 Analysis experiment

Training Format and Shortcut Learning Analysis. To verify that performance improvements stem from genuine spatial reasoning rather than shortcut learning, we conducted controlled experiments comparing multiple-choice and direct answer training formats on the same dataset described in Section 3.3. As shown in Table 3, both formats yielded substantial improvements over the baseline, with multiple-choice achieving 82.09% and direct answer achieving 79.34% overall accuracy. The negligible performance gap confirms that MVSM's gains reflect enhanced spatial understanding rather than exploitation of multiple-choice structural patterns, validating robust generalization across response formats.

Multi-Backbone Generalization Validation. To demonstrate the broader applicability of our methodology, we evaluated MVSM across multiple VLM architectures using identical training configurations (Section 5.1). As shown in Table 3, all three backbones achieved consistent and substantial improvements: Qwen2.5-VL(3B) from 35.85% to 82.09% (+46.24%), InternVL-2B from 34.98% to 76.45% (+41.47%), and Qwen2.5-VL(7B) from 36.85% to 83.01% (+46.16%). The uniform gains across different model families and parameter scales confirm the architecture-agnostic effectiveness of our perspective-aware training approach.

Analysis and Discussion. The results above demonstrate that targeted fine-tuning with task-specific data consistently yields substantial spatial reasoning improvements. This pattern is corroborated across multiple independent benchmarks: Multi-SpatialMLLM [37] achieves an average gain of 35.68% on MultiSPA [37], with certain subtasks improving by up to 71.99%; targeted fine-tuning of MiniGPT-v2 [4] on EmbSpatial-Bench [11] yields +34.25%, and analogous trends are observed in MindCube [42] and Cambrian-S [40].

Although each of these benchmarks—including ViewSpatial-Bench—addresses a distinct spatial problem and serves its own unique research purpose, they collectively exhibit the same pronounced improvement pattern under targeted fine-tuning. This consistency across diverse tasks and model families suggests that current VLMs still possess considerable untapped capacity for spatial understanding, and that spatial reasoning represents a systematic, cross-domain weakness remediable through dedicated supervision. These converging findings underscore the generality and importance of continued research in this direction.

5.4 Empowering Spatial Interaction Application

To further validate MVSM's spatial understanding capabilities in practical applications, we evaluated its performance on VSI-Bench [39] in typical tasks requiring perspective transformation, including Object Relative Direction and Route Planning subtasks. Additionally, we constructed a small application evaluation dataset, ViewSpatial Interaction Application Dataset (VSI-App), encompassing both indoor and outdoor scenarios, specifically designed to assess spatial orientation recognition abilities in embodied interaction environments, with particular focus on the requirements for dynamic scene and multi-perspective understanding during human-machine interaction.

Table 3: Comprehensive analysis of MVSM training robustness across different question formats and model architectures. MC denotes Multiple Choice format, DA denotes Direct Answer format.

Experiment	Backbone Model	Original	MVSM	Improvement
Training Format	Qwen2.5-VL (3B)	35.85%	**82.09**% *(MC)*	+46.24%
			79.34% *(DA)*	+43.49%
Multi-Backbone	Qwen2.5-VL (3B)	35.85%	**82.09%**	+46.24%
	Qwen2.5-VL (7B)	36.85%	**83.01%**	+46.16%
	InternVL2.5 (2B)	34.98%	**76.45%**	+41.47%

Transfer Learning Performance As shown in Table 4, we assessed MVSM's generalization capabilities on both VSI-Bench and our custom VSI-App benchmark. The specific construction process and evaluation methods of the VSI-App are shown in Appendix B.4.

Table 4: Performance comparison of our Multi-View Spatial Model against its backbone.

	VSI-Bench			VSI-App		
Model	Rel Dir	Route Plan	Average	Indoor	Outdoor	Average
GPT-4o [15]	41.30	31.50	39.66	34.00	27.00	30.50
Qwen2.5-VL(3B) [1]	46.00	21.90	41.97	18.00	27.00	22.50
MVSM	**46.93** ↑ 0.93	**31.44** ↑ 9.54	**44.34** ↑ 2.37	**41.00** ↑ 23.00	**36.00** ↑ 9.00	**38.50** ↑ 16.00

VSI-Bench Evaluation: We selected two representative tasks requiring perspective transformation abilities: Object Relative Direction and Route Planning. The former requires determining precise spatial relationships between objects in complex indoor scenes, while the latter involves inferring and completing coherent navigation paths through 3D environments. As shown in Table 4, MVSM outperforms its backbone model on both tasks, with particularly significant gains in Route Planning (+9.54%).

Notably, this improvement emerges without any explicit route planning supervision in our training data, suggesting that perspective-aware fine-tuning cultivates a more general spatial representation that extends naturally from static object-level relationships to dynamic trajectory reasoning. These results indicate that multi-viewpoint spatial training implicitly enriches VLMs' understanding of 3D scene geometry, enabling more reliable spatial inference across diverse task formulations beyond those seen during training.

VSI-App Evaluation: To further approximate real-world interaction scenarios, we constructed VSI-App, a specialized evaluation dataset of 50 scenes

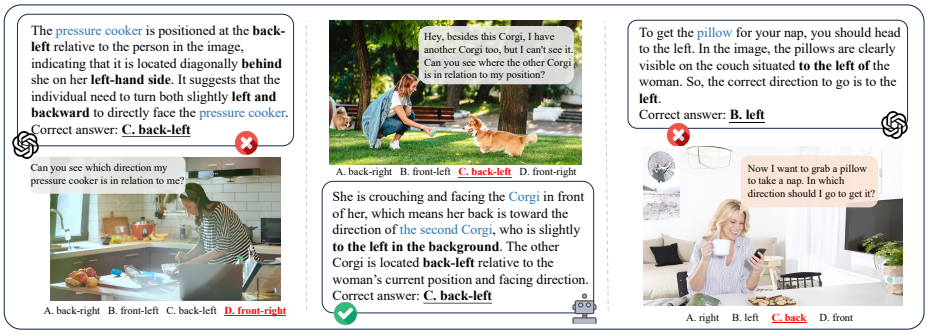

Fig. 4: The image compares spatial reasoning performance between GPT-4o and MVSM on the VSI-App dataset, showing several examples where MVSM correctly answers perspective-taking questions about object locations, while GPT-4o makes errors when attempting to determine spatial relationships from another person's viewpoint.

(25 indoor, 25 outdoor) designed to assess human-centric spatial reasoning in embodied contexts. The benchmark requires models to perform spatial reasoning from human first-person perspectives, generating responses that conform to human cognitive patterns. MVSM shows substantial improvement in indoor environments (+20.00%) and modest gains in outdoor scenarios (+4.00%). This performance pattern reveals an interesting domain gap: indoor environments with structured spatial relationships better align with our training distribution, while outdoor scenes pose greater challenges despite still showing improvement.

Perspective Confusion Analysis The performance improvement on our benchmarks stems directly from MVSM's enhanced ability to maintain consistent perspective representations. To illustrate this capability, Figure 4 contrasts MVSM with GPT-4o on representative VSI-App examples requiring perspective transformation. While GPT-4o demonstrates some ability to locate objects from human perspectives, it frequently defaults to camera-centric judgments for orientation determinations, resulting in perspective confusion.

Analysis of failure modes reveals that models without perspective-aware training demonstrate inconsistent spatial judgments within single responses, alternating between human and camera perspectives. This suggests they lack a coherent internal model of 3D space that can be navigated from different viewpoints. In contrast, MVSM maintains consistent adherence to the specified perspective frame, even in challenging cases requiring multiple spatial transformations.

6 Conclusions

In this work, we present ViewSpatial-Bench, a dedicated benchmark for evaluating multi-perspective spatial localization capabilities of vision-language models

across five distinct task types. Our assessment of various advanced VLMs reveals significant limitations in their spatial reasoning abilities. By developing an automated spatial annotation pipeline and constructing a large-scale multi-perspective dataset, we successfully trained our Multi-View Spatial Model (MVSM), which achieves substantial overall performance improvements on ViewSpatial-Bench tasks. Further experiments on VSI-Bench and our custom VSI-App dataset demonstrate MVSM's generalization capabilities to real-world embodied interaction scenarios. Our work establishes a foundation for spatially intelligent VLMs that better align with human cognitive patterns in embodied environments, representing an important step toward more intuitive and effective human-machine spatial communication.

Acknowledgements

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123100), National Natural Science Foundation of China (No. 62506332), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (NO. 2026C02A1223) and ZJU Kunpeng&Ascend Center of Excellence.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al.: An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247 (2024)
3. 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)
4. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. arxiv 2023. arXiv preprint arXiv:2310.09478 (2023)
5. 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)
6. 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. arXiv preprint arXiv:2406.01584 (2024)
7. Cho, Y., Yu, H., Kang, S.J.: Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. IEEE Transactions on Multimedia **26**, 5823–5833 (2023)

8. 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)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
10. Deng, H., Zou, D., Ma, R., Luo, H., Cao, Y., Kang, Y.: Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. arXiv preprint arXiv:2503.07065 (2025)
11. Du, M., Wu, B., Li, Z., Huang, X., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. arXiv preprint arXiv:2406.05756 (2024)
12. Feng, Z., Xue, R., Yuan, L., Yu, Y., Ding, N., Liu, M., Gao, B., Sun, J., Wang, G.: Multi-agent embodied ai: Advances and future directions. arXiv preprint arXiv:2505.05108 (2025)
13. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
14. 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)
15. 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)
16. Kamath, A., Hessel, J., Chang, K.W.: What's" up" with vision-language models? investigating their struggle with spatial reasoning. arXiv preprint arXiv:2310.19785 (2023)
17. 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. arXiv preprint arXiv:2504.17207 (2025)
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
20. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3707–3717 (2025)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
22. Liu, Z., Liu, M., Chen, J., Xu, J., Cui, B., He, C., Zhang, W.: Fusion: Fully integration of vision-language representations for deep cross-modal understanding. arXiv preprint arXiv:2504.09925 (2025)
23. Ma, C., Lu, K., Cheng, T.Y., Trigoni, N., Markham, A.: Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and

interacting 3d priors. Advances in neural information processing systems **37**, 68803–68832 (2024)
24. Ma, W., Chen, H., Zhang, G., de Melo, C.M., Chen, J., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. arXiv preprint arXiv:2412.07825 (2024)
25. OpenAI: GPT-5 System Card. Technical report, OpenAI (Aug 2025), accessed: 2025-08-10
26. Paz-Argaman, T., Palowitch, J., Kulkarni, S., Baldridge, J., Tsarfaty, R.: Where do we go from here? multi-scale allocentric relational inferencefrom natural spatial descriptions. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1026–1040 (2024)
27. Rajabi, N., Kosecka, J.: Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. arXiv preprint arXiv:2406.13246 (2024)
28. Seed, B.: Seed1. 8 model card: Towards generalized real-world agency. Tech. rep., Technical report (model card), December 2025. URL https://lf3-static ... (2025)
29. Shiri, F., Guo, X.Y., Far, M.G., Yu, X., Haffari, G., Li, Y.F.: An empirical analysis on spatial reasoning capabilities of large multimodal models. arXiv preprint arXiv:2411.06048 (2024)
30. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. arXiv preprint arXiv:2411.16537 (2024)
31. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
32. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
33. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
34. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
35. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), `https://qwen.ai/blog?id=qwen3.5`
36. Wang, Z., Zhang, Z., Pang, T., Du, C., Zhao, H., Zhao, Z.: Orient anything: Learning robust object orientation estimation from rendering 3d models. arXiv preprint arXiv:2412.18605 (2024)
37. Xu, R., Wang, W., Tang, H., Chen, X., Wang, X., Chu, F.J., Lin, D., Feiszli, M., Liang, K.J.: Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. arXiv preprint arXiv:2505.17015 (2025)
38. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
39. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
40. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)

41. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
42. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
43. Zang, Y., Yun, T., Tan, H., Bui, T., Sun, C.: Pre-trained vision-language models learn discoverable visual concepts. arXiv preprint arXiv:2404.12652 (2024)
44. Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X.: How to enable llm with 3d capacity? a survey of spatial reasoning in llm. arXiv preprint arXiv:2504.05786 (2025)
45. Zhang, W., Ng, W.E., Ma, L., Wang, Y., Zhao, J., Koenecke, A., Li, B., Wang, L.: Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. arXiv preprint arXiv:2412.12693 (2024)
46. Zhang, Y., Li, B., Liu, H., Lee, Y., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (2024)
47. Zhao, X., Cai, W., Tang, L., Wang, T.: Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. arXiv preprint arXiv:2410.09874 (2024)
48. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8995–9006 (2025)
49. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., 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)