

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

Dingming Li<sup>1,2,\*</sup> Hongxing Li<sup>1,\*</sup> Zixuan Wang<sup>1</sup> Yuchen Yan<sup>1</sup>

Hang Zhang<sup>1</sup> Siqi Chen<sup>1</sup> Guiyang Hou<sup>1</sup> Shengpei Jiang<sup>3</sup>

Wenqi Zhang<sup>1</sup> Yongliang Shen<sup>1,†</sup> Weiming Lu<sup>1</sup> Yueting Zhuang<sup>1</sup>

<sup>1</sup> Zhejiang University <sup>2</sup> University of Electronic Science and Technology of China

<sup>3</sup> The Chinese University of Hong Kong

lidinm@std.uestc.edu.cn, sheny1@zju.edu.cn

Project: <https://zju-real.github.io/ViewSpatial-Page>

## 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, the first comprehensive benchmark designed specifically for multi-viewpoint spatial localization recognition evaluation across five distinct task types, 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, highlighting the efficacy of our approach. 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.

## 1 Introduction

While Vision-Language Models (VLMs) demonstrate remarkable capabilities in visual content understanding and reasoning [1, 2, 3], they exhibit significant limitations when confronted with complex tasks requiring cross-viewpoint comprehension and spatial reasoning [4, 5]. Specifically, current VLMs perform adequately in egocentric spatial judgments but struggle to interpret and reason about spatial relationships from alternative entity perspectives [6]. 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

\* The first two authors have equal contributions. This work was done when the first author was an intern at Zhejiang University.

† Corresponding author.



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

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 [7].

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 [8], and multi-agents collaboration [9]. 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 [10].

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 [11]. 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 [2, 12]. 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, the first comprehensive benchmark for evaluating spatial localization from both camera and human perspectives. 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*, the first comprehensive 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. Through systematic evaluation, we identify fundamental limitations in current models’ perspective-based spatial reasoning, particularly in 3D embodied environments. Our model achieves 46.24% improvement over baselines, demonstrating our methodology’s effectiveness.

## 2 Related Works

**Spatial Reasoning with VLMs.** Recently, VLMs have demonstrated significant advancements in understanding and reasoning about visual content [13, 14, 15]. Both proprietary and open-source models have achieved impressive performance in visual question answering (VQA), image captioning, and complex multimodal reasoning tasks. These models typically employ Transformer [16] architectures, incorporating image encoders and vision-language fusion modules [17, 18, 19], and are pre-trained on large-scale image-text pairs to establish strong connections between visual elements and linguistic descriptions [20].

However, despite current VLMs’ exceptional performance on certain visual reasoning tasks, their spatial understanding capabilities remain fundamentally limited [2, 21]. 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 [4, 22]. This constraint impairs their generalization capabilities and practical utility in tasks requiring perspective transformation or third-person spatial comprehension. Consequently, developing models with stronger perspective-taking awareness has emerged as a critical challenge for advancing multimodal intelligence to higher levels of sophistication.

**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 [23] and What’sUP [24] concentrate on assessing models’ abilities to recognize spatial relationships between objects in two-dimensional images, while VSI-Bench [21] 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 [3].

Although some benchmarks have begun to address cross-viewpoint spatial understanding, such as 3DSRBench [25] and SPHERE [26], which evaluate existing VLMs across various spatial tasks, 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 [27] and MS-CoCo [28]. 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. Oir.), 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. Oir.), 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.

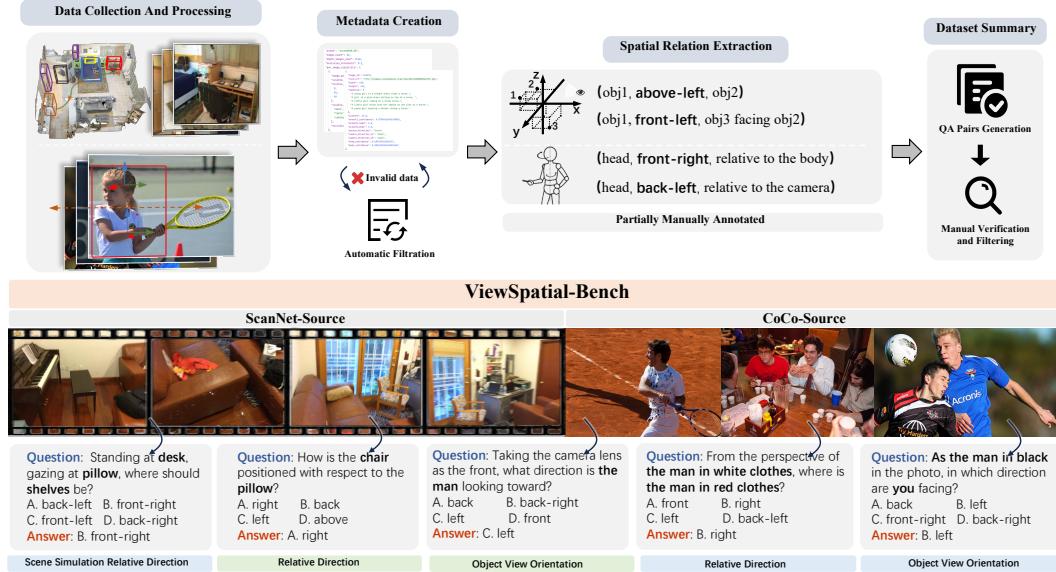


Figure 2: ViewSpatial-Bench construction pipeline. From data collection to QA generation across camera perspective (blue) and human perspective (green) tasks. The pipeline includes metadata creation, automatic filtering, spatial relation extraction, and manual verification.

### 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 human-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 [29]) 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. Oir. and Per-Obj. Oir. 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.

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 [30] 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.

---

**Algorithm 1** Maximum Coverage Sampling

---

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

- 1: Initialize  $S \leftarrow \emptyset$
- 2: Initialize  $U \leftarrow \emptyset$  {Covered voxels set}
- 3: **while** size of  $S$  is less than  $K$  **do**
- 4:     Select  $f^* = \arg \max_{f_k \in F \setminus S} |V_k \setminus U|$
- 5:     Add  $f^*$  to  $S$
- 6:     Update  $U \leftarrow U \cup V_{f^*}$
- 7:     **if** Stop condition is met **then**
- 8:         **break**
- 9:     **end if**
- 10: **end while**
- 11: **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

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

---

**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.,  $22.5^\circ$  to  $67.5^\circ$  as "front-right,"  $67.5^\circ$  to  $112.5^\circ$  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:  $63.43^\circ$  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 [22, 23]. More dataset details are in Appendix B.

### 3.3 Dataset Statistics

Figure 3 illustrates the five task categories in ViewSpatial-Bench and their respective proportions. To ensure balanced evaluation across viewpoints, we constructed approximately equal amounts of data for camera-perspective (48.4%) and human-perspective (51.6%) tasks. This balanced distribution enables fair comparison of spatial reasoning capabilities from different observational frameworks. For the Relative Direction recognition

task from camera viewpoints, which more directly demonstrates 3D scene understanding, we developed additional data to enrich spatial information diversity.

Figure 4 shows the frequency distribution of spatial prepositions and objects in ViewSpatial-Bench. As illustrated in Figure 4(a), our benchmark incorporates a comprehensive set of directional terms, with balanced representation of primary directions ("front", "back", "left", "right") and compound directions ("front-left", "back-right", etc.). This diverse coverage ensures thorough evaluation of VLMs' ability to process complex spatial relationships from multiple perspectives. Figure 4(b) depicts the distribution of the top 20 objects in ViewSpatial-Bench. The object distribution reflects common entities encountered in everyday environments, with furniture items (chair, table, sofa, desk) and personal objects (person, cup, bottle) well represented. This ensures practical relevance of the benchmark to real-world spatial reasoning scenarios, particularly for embodied AI applications that must navigate and interact with common objects.

## 4 Multi-View Spatial Model

To address the limitations in perspective-dependent spatial reasoning identified in current VLMs, we developed the Multi-View Spatial Model (MVSM) through a systematic enhancement approach. Our methodology combines high-quality training data with a specialized fine-tuning strategy designed specifically for multi-viewpoint spatial understanding. 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 [27] and MS-COCO [28] training sets, supplemented with Spatial-MM [4] data for the Per-Rel. Dir. task where full automation proved challenging due to complex human spatial coordinates and environmental contexts. Our Multi-Perspective Fine-Tuning strategy explicitly trains the model to reason from different observational viewpoints, enabling MVSM to develop a

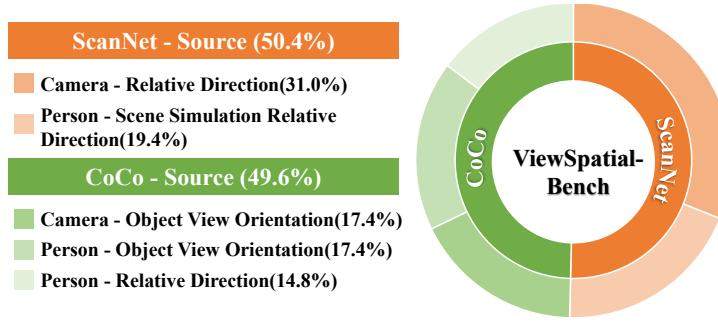


Figure 3: Distribution of task categories in ViewSpatial-Bench, balanced between ScanNet-Source and CoCo-Source approaches, with five distinct subtasks for comprehensive evaluation of spatial reasoning across different viewpoints.

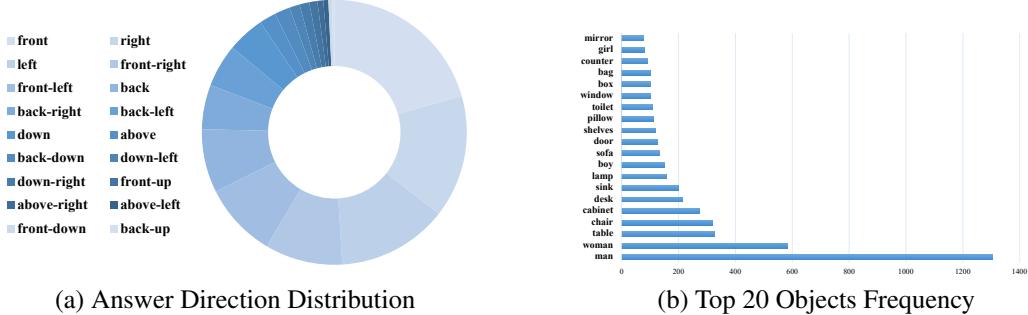


Figure 4: Frequency distributions in ViewSpatial-Bench. (a) Distribution of spatial prepositions, showing comprehensive coverage of directional relationships. (b) Frequency of the top 20 objects, demonstrating the benchmark’s focus on common entities encountered in everyday environments.

more unified representation of 3D spatial relationships that supports robust reasoning across both camera and human perspectives.

## 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 [31, 32], LLaVA-NeXT-Video [33], LLaVA-OneVision [34], Llama-3.2-Vision [35], Kimi-VL-Instruct [36], and Qwen2.5-VL [37]; (2) Proprietary models: GPT-4o [38] and Gemini-2.0-Flash [39].

**Implementation Details.** For our fine-tuning experiments, we use Qwen2.5-VL [37] as the backbone model. Following standard practice in efficient adaptation, we freeze the vision encoder and multi-modal projector while keeping the language model trainable. The model is trained for 3 epoch with an effective batch size of 16, achieved through gradient accumulation (4 steps with per-device batch size of 1) across 4 NVIDIA A100 (40GB) GPUs. The entire training process requires approximately 8.5 GPU hours, making our approach computationally efficient and accessible. More model training and evaluation details are shown in Appendix C.

### 5.2 Main Results

As shown in Table 1, 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%) and Gemini-2.0-Flash (32.56%) 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 this is not merely an implementation issue but a systematic deficiency in how these models conceptualize spatial relationships.

**Egocentric vs. allocentric reasoning gap:** Most VLMs exhibit an intriguing pattern wherein their spatial localization accuracy from camera perspectives (averaging 33.2%) falls below their performance from human viewpoints (averaging 35.7%). This contradicts the intuitive expectation that egocentric perspective (camera-based) should be easier than allocentric reasoning (human-based). This finding aligns with our observation that VLMs lack the perspective-taking ability that humans naturally possess, and suggests that current vision-language architectures may implicitly encode certain spatial biases that favor third-person viewpoints, potentially due to the prevalence of such compositions in web-harvested training data.

Model	Camera-based Tasks			Person-based Tasks			Overall	
	Rel. Dir.	Obj. Ori.	Avg.	Obj. Ori.	Rel. Dir.	Sce. Sim.		
InternVL2.5 (2B)[31]	38.52	22.59	32.79	47.09	40.02	25.70	37.04	34.98
Qwen2.5-VL (7B)[37]	46.64	29.72	40.56	37.05	35.04	28.78	33.37	36.85
LLaVA-NeXT-Video (7B)[33]	26.34	19.28	23.80	44.68	38.60	29.05	37.07	30.64
LLaVA-OneVision (7B)[34]	29.84	26.10	28.49	22.39	31.00	26.88	26.54	27.49
InternVL2.5 (8B)[31]	49.41	<b>41.27</b>	46.48	46.79	42.04	<b>32.85</b>	40.20	<b>43.24</b>
Llama-3.2-Vision (11B)[35]	25.27	20.98	23.73	51.20	32.19	18.82	33.61	28.82
InternVL3 (14B)[32]	<b>54.65</b>	33.63	<b>47.09</b>	33.43	37.05	31.86	33.88	40.28
Kimi-VL-Instruct (16B)[36]	26.85	22.09	25.14	<b>63.05</b>	<b>43.94</b>	20.27	<b>41.52</b>	33.58
GPT-4o[38]	41.46	19.58	33.57	42.97	40.86	26.79	36.29	34.98
Gemini 2.0 Flash[39]	45.29	12.95	33.66	41.16	32.78	21.90	31.53	32.56
Qwen2.5-VL (3B)[37] [Backbone]	43.43	33.33	39.80	39.16	28.62	28.51	32.14	35.85
<b>Multi-View Spatial Model</b>	<b>83.59</b>	<b>87.65</b>	<b>85.05</b>	<b>90.16</b>	<b>71.14</b>	<b>75.75</b>	<b>79.31</b>	<b>82.09</b>
<i>Improvement over backbone</i>	+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

Table 1: Zero-shot performance on ViewSpatial-Bench. Accuracy comparison across multiple VLMs on camera and human perspective spatial tasks. Our Multi-View Spatial Model (MVSM) significantly outperforms all baseline models across all task categories, demonstrating the effectiveness of our multi-perspective spatial fine-tuning approach.

**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, yet show the opposite pattern for human perspective tasks (42.6% for Object View Orientation vs. 36.9% for 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 human-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 Empowering Spatial Interaction Application

To further validate MVSM’s spatial understanding capabilities in practical applications, we evaluated its performance on VSI-Bench [22] 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.

#### 5.3.1 Transfer Learning Performance

As shown in Table 2, 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.

**VSI-Bench Evaluation:** We selected two representative tasks requiring perspective transformation abilities: Object Relative Direction and Route Planning. The former requires determining spatial relationships between objects in complex indoor scenes, while the latter involves inferring and completing reasonable navigation paths. MVSM outperforms its backbone model in both tasks, with particularly significant gains in Route Planning (+9.54%). This improvement demonstrates MVSM’s

Model	VSI-Bench			VSI-App		
	Rel Dir	Route Plan	Average	Indoor	Outdoor	Average
Qwen2.5-VL(3B)[37]	46.00	21.90	41.97	24.00	20.00	22.00
<b>MVSM</b>	<b>46.93</b> ↑ 0.93	<b>31.44</b> ↑ 9.54	<b>44.34</b> ↑ 2.37	<b>44.00</b> ↑ 20.00	<b>24.00</b> ↑ 4.00	<b>34.00</b> ↑ 12.00

Table 2: Performance comparison of our Multi-View Spatial Model against its backbone.

enhanced ability to model not just static spatial relationships but also dynamic trajectories through 3D environments, which emerged from our perspective-aware training approach without explicit route planning optimization.

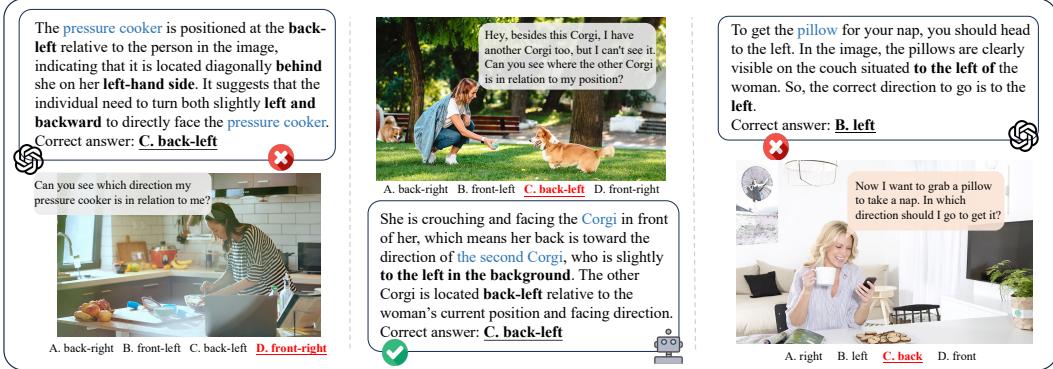


Figure 5: The image compares spatial reasoning performance between GPT-4o and MVSM on the VSI-App dataset, showing several examples where MVSM correctly answers perspective-taking questions about object locations, while GPT-4o makes errors when attempting to determine spatial relationships from another person’s viewpoint.

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

### 5.3.2 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 5 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, the first comprehensive 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.

## References

- [1] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.
- [2] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024.
- [3] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *arXiv preprint arXiv:2411.16537*, 2024.
- [4] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*, 2024.
- [5] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.
- [6] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025.
- [7] Tzuf Paz-Argaman, John Palowitch, Sayali Kulkarni, Jason Baldridge, and Reut Tsarfaty. Where do we go from here? multi-scale allocentric relational inference from natural spatial descriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1026–1040, 2024.
- [8] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2410.09874*, 2024.
- [9] Zhaohan Feng, Ruiqi Xue, Lei Yuan, Yang Yu, Ning Ding, Meiqin Liu, Bingzhao Gao, Jian Sun, and Gang Wang. Multi-agent embodied ai: Advances and future directions. *arXiv preprint arXiv:2505.05108*, 2025.
- [10] Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. *arXiv preprint arXiv:2412.04383*, 2024.
- [11] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *arXiv preprint arXiv:2403.13438*, 2024.
- [12] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.
- [13] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.
- [14] Harsh Lunia. Can vlms be used on videos for action recognition? llms are visual reasoning coordinators. *arXiv preprint arXiv:2407.14834*, 2024.
- [15] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [17] Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE Transactions on Multimedia*, 26: 5823–5833, 2023.
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [19] Zheng Liu, Mengjie Liu, Jingzhou Chen, Jingwei Xu, Bin Cui, Conghui He, and Wentao Zhang. Fusion: Fully integration of vision-language representations for deep cross-modal understanding. *arXiv preprint arXiv:2504.09925*, 2025.
- [20] Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts. *arXiv preprint arXiv:2404.12652*, 2024.
- [21] ENCE UNDER AMBIGUITIES. Do vision-language models represent space and how? evaluating spatial frame of refer, 2025. URL <https://arxiv.org/abs/2410.17385>.
- [22] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.
- [23] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024.
- [24] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's " up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.
- [25] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.
- [26] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Allison Koencke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. *arXiv preprint arXiv:2412.12693*, 2024.
- [27] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [29] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. *arXiv preprint arXiv:2412.00493*, 2024.
- [30] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*, 2024.
- [31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [32] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [33] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.
- [34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

- [35] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [36] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [37] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [38] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [39] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

## A Limitations

While ViewSpatial-Bench represents a significant step forward in evaluating multi-perspective spatial reasoning in VLMs, several limitations merit acknowledgment:

**Annotation Challenges for Human-Perspective Tasks.** The Person-perspective Relative Direction task presented substantial annotation challenges. The inherent complexity of human spatial coordinates and environmental contexts in natural images prevented full automation of the annotation process. This necessitated manual labeling, which introduces both scaling constraints and potential annotator biases. Future work could explore semi-supervised approaches that might reduce the reliance on manual annotation while maintaining data quality.

**Domain Constraints in Environmental Coverage.** Our Camera-perspective Relative Direction tasks utilize exclusively indoor environments from ScanNet, potentially limiting generalizability to outdoor settings. As our transfer learning experiments on VSI-App suggest, there exists a substantial domain gap between indoor and outdoor spatial reasoning tasks. Outdoor environments present different spatial scales, object densities, and visual characteristics that may require specialized training approaches beyond those presented in this work.

**Static vs. Dynamic Spatial Reasoning.** ViewSpatial-Bench evaluates only static spatial orientation comprehension without addressing dynamic spatial reasoning scenarios where objects or observers move through environments. Such dynamic reasoning represents an important aspect of embodied spatial cognition relevant to many practical applications, including robot navigation and interactive systems. Extending our benchmark to incorporate temporal sequences and motion-based spatial reasoning would provide a more comprehensive evaluation framework for embodied AI systems.

These limitations point to promising directions for future research that could build upon the foundation established by ViewSpatial-Bench while addressing its current constraints.

## B Data Details

## B.1 Dataset Collection and Unification

**ScanNet Data Collection.** We employ a three-stage video frame sampling strategy to optimize benchmark data quality: first extracting all video frames, then uniformly sampling every 10th frame, and finally applying maximum frame sampling to select the minimal yet comprehensive set of consecutive frames that capture complete scene information.

For 3D bounding box visibility analysis, we utilize a depth-aware projection technique that transforms 3D bounding boxes from world coordinates to camera view while accounting for occlusions. Our implementation aligns depth and color frames using scale factors (1000.0 mm to m) and handles resolution differences through proportional coordinate mapping. The occlusion detection compares the computed depth of 3D bounding box vertices against the measured depth from sensor data with a 0.1m threshold, enabling accurate determination of vertex visibility. This approach generates precise visibility annotations by requiring at least 1% of vertices to be visible for an object to be considered present in a frame, enhancing the fidelity of our object detection and 3D reasoning benchmarks.

**MS-COCO Data Collection.** Based on MS-COCO dataset annotations, we filter samples containing biological objects that occupy at least 20% of the image area to ensure sufficient visual salience of target objects. We subsequently employ manual annotation to filter out samples where gaze direction significantly deviates from head orientation, ensuring consistency in spatial orientation labeling. The filtered samples are then processed by the Orient-Anything-Large model for automatic head and body orientation angle annotation. Given that this model exhibits labeling errors when processing low-resolution images or objects with ambiguous directional tendencies, we conduct focused manual verification and data correction on extreme angle samples (excessively large or small angles). This quality assurance mechanism ensures the annotation accuracy of the final dataset.

**QA Pair Generation.** We extract object information and corresponding angle annotations from metadata for each sample. Object names are filled into predefined question templates, with computed angles serving as ground truth answers to construct multiple-choice questions. The question templates used are detailed in Table 3.



Figure 6: Wordcloud of object categories.

Task	Question Template
Cam-Rel. Dir.	<ul style="list-style-type: none"> <li>• Can you describe the position of the {object1} relative to the {object2}?</li> <li>• Could you tell me the location of the {object1} in comparison to the {object2}?</li> <li>• Where is the {object1} in relation to the {object2}?</li> <li>• Where is the {object1} located compared to the {object2} from the camera’s perspective?</li> <li>• How is the {object1} positioned with respect to the {object2}?</li> <li>• If you’re looking at the {object2}, where would you find the {object1}?</li> </ul>
Cam-Obj. Dir.	<ul style="list-style-type: none"> <li>• With the camera’s viewpoint as the front, which direction is {object} facing in the image?</li> <li>• Taking the camera lens as the front, what direction is {object} looking toward?</li> <li>• Taking the camera’s viewpoint as the front, which way is {object} facing in the image?</li> <li>• Considering the camera’s perspective as the front, what direction is {object} facing within the picture?</li> </ul>
Per-Obj. Dir.	<ul style="list-style-type: none"> <li>• Imagine you’re {object} in this image — which direction are you facing?</li> <li>• Suppose you are in {object}’s position, what direction are you facing?</li> <li>• Picture yourself as {object}; which way are you looking in the scene?</li> <li>• As {object} in the photo, in which direction are you facing?</li> </ul>
Per-Sce. Sim.	<ul style="list-style-type: none"> <li>• Imagine standing at {object1} looking towards {object2}, where is {object3}?</li> <li>• When positioned at {object1} facing {object2}, where can you find {object3}?</li> <li>• If you stand at {object1} facing {object2}, where is {object3}?</li> <li>• Standing at {object1}, gazing at {object2}, where should {object3} be?</li> </ul>

Table 3: Prompt templates used to generate spatial reasoning questions across four tasks. Object names are inserted into the templates to form natural language questions, which are later paired with direction-based multiple-choice answers derived from scene metadata.

## B.2 Data Statistics

As shown in the word cloud analysis in Figure 6, our dataset is primarily constructed around two major categories: humans and objects, which aligns with our dual spatial localization task design targeting both camera and human perspectives. Tabl 4 provides a detailed breakdown of sample distributions across different task types in ViewSpatial-Bench.

## B.3 Data Cases

Figures 7–9 illustrates response examples from different models across various question types in ViewSpatial-Bench.

## B.4 VSI-App Dataset Construction

For the ViewSpatial Interaction Application Dataset (VSI-App), we employ a three-stage human curation approach to construct a dataset specifically designed to evaluate multi-view spatial models (MVSM) capabilities in spatial reasoning for human-computer interaction under Out-of-Distribution scenarios. Initially, two professional annotators carefully screened and downloaded 50 high-quality scene images from professional online image platforms, with 25 indoor and 25 outdoor scenes respectively. Image selection strictly adheres to the following criteria: scenes must be highly consistent with indoor/outdoor themes, contain rich three-dimensional spatial hierarchical information, include clearly identifiable human subjects as viewpoint references, and demonstrate explicit spatial relationships and potential interaction possibilities between humans and other objects in the scene. This meticulous scene selection ensures that the dataset can adequately simulate the complex spatial environments of real-world human-computer interactions.

In the question annotation phase, two annotators conduct in-depth spatial analysis of the primary human subjects in each image, focusing on two core interaction scenarios: first, spatial cognition questions where human subjects inquire about the relative positions of other objects from their first-person perspective, and second, path planning and navigation orientation questions from the human’s current position to target locations. The annotators completely abandon template-based QA generation methods, directly employing natural language that closely resembles daily communication for question descriptions, while meticulously designing accurate ground truth answers and plausible distractors for each question. This natural language annotation approach not only enhances question diversity and authenticity, but more importantly captures the linguistic expression habits and cognitive patterns of humans in actual spatial interactions.

VSI-App aims to verify whether MVSM can accurately understand and respond to spatial reasoning inquiries from human perspectives when confronted with realistic human-computer interaction scenarios, thereby evaluating the model’s generalization capability and practical utility. Evaluation follows a multiple-choice format, with specific examples shown in Figure 5.

	Camera			Person			Overall
	Rel. Dir.	Obj. Dir.	Sum.	Obj. Dir.	Rel. Dir.	Sce. Sim.	
Test	1773	996	2769	996	842	1105	2943 <b>5712</b>
Train	13644	8954	22598	8954	1014	10309	20277 <b>42875</b>

Table 4: Sample counts for different tasks in ViewSpatial-Bench evaluation and MVSM training data.

## C Experiments

### C.1 Implementation Details

We select Qwen2.5-VL as the base model for supervised fine-tuning. The Cam-Rel. Dir., Cam-Obj. Dir., Per-Obj. Dir., and Per-Sce. Sim. tasks in the training dataset are generated through our automated construction pipeline using unified QA templates. The Per-Rel. Dir. task is constructed based on the Spatial-MM dataset, with language models employed to polish questions and enhance sample diversity. The distribution of training samples across tasks is detailed in Table 4.

### C.2 Evaluation Details

**ViewSpatial-Bench evaluation.** We evaluate all models under zero-shot settings, where models must directly predict the correct option based on given images and questions. Accuracy is calculated by comparing model predictions with ground truth answers. The prompt template used for evaluation is shown below.

Zero-shot Evaluation Prompt
Question:{question} Choices:{choices} Reply only to the corresponding option. Answer:

**VSI-Bench evaluation.** We follow the original paper’s experimental settings for VSI-Bench evaluation. We employ the lmms-eval framework to conduct zero-shot testing with a batch size of 1 and maximum frame count set to 32. All models are evaluated on a single GPU environment (A6000 48G) using the accelerate launcher.

**VSI-App dataset evaluation.** Since VSI-App is a small-scale test benchmark designed for Out-of-Distribution scenarios, we adopt a repeated testing strategy to enhance evaluation reliability. Specifically, we generate 5 different option orderings for each question sample and conduct 5 independent tests for each model on these reordered samples. The final answer is determined through a voting mechanism, selecting the option with the highest frequency across the 5 tests for the same question as the prediction result. This method effectively reduces the potential impact of option ordering on model predictions.

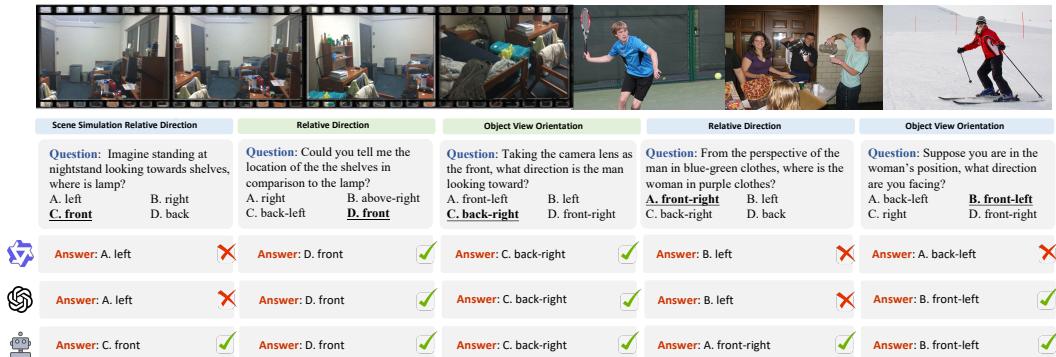


Figure 7: ViewSpatial-Bench Examples (Part1). Performance comparison of three models (Qwen2.5-VL(3B), GPT-4o, and MVSM) on five spatial reasoning tasks from camera perspective (blue) and human perspective (green).

Scene Simulation Relative Direction	Relative Direction	Object View Orientation	Relative Direction	Object View Orientation
	<b>Question:</b> When positioned at refrigerator facing counter, where can you find window? A. back-left B. front-left C. front-right D. back-right	<b>Question:</b> How is the refrigerator positioned with respect to the window? A. front-up B. right C. above D. back-up	<b>Question:</b> Taking the camera's viewpoint as the front, which way is the girl facing in the image? A. front B. left C. right D. back	<b>Question:</b> From the perspective of the blonde woman, where is the man in black positioned? A. left B. front-right C. front D. right
	<b>Answer:</b> B. front-left		<b>Answer:</b> B. right	
	<b>Answer:</b> B. front-left		<b>Answer:</b> B. right	
	<b>Answer:</b> D. back-right		<b>Answer:</b> B. right	
	<b>Question:</b> When positioned at refrigerator facing counter, where can you find window? A. back-left B. front-left C. front-right D. back-right	<b>Question:</b> How is the refrigerator positioned with respect to the window? A. front-up B. right C. above D. back-up	<b>Question:</b> Taking the camera's viewpoint as the front, which way is the girl facing in the image? A. front B. left C. right D. back	<b>Question:</b> From the perspective of the blonde woman, where is the man in black positioned? A. left B. front-right C. front D. right
	<b>Answer:</b> B. front-left		<b>Answer:</b> B. right	
	<b>Answer:</b> B. front-left		<b>Answer:</b> B. right	
	<b>Answer:</b> D. back-right		<b>Answer:</b> B. right	

Figure 8: ViewSpatial-Bench Examples (Part2).

Scene Simulation Relative Direction	Relative Direction	Object View Orientation	Relative Direction	Object View Orientation
	<b>Question:</b> Standing at table, gazing at chair, where should books be? A. front B. back-left C. right D. back	<b>Question:</b> If you're looking at the chair, where would you find the table? A. back-down B. back-up C. right D. back-left	<b>Question:</b> With the camera's viewpoint as the front, which direction is the man facing in the image? A. front B. right C. back D. left	<b>Question:</b> From the perspective of the woman in blue clothes, where is the woman in white clothes located? A. left B. right C. front D. front-right
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	
	<b>Question:</b> Standing at table, gazing at chair, where should books be? A. front B. back-left C. right D. back	<b>Question:</b> If you're looking at the chair, where would you find the table? A. back-down B. back-up C. right D. back-left	<b>Question:</b> With the camera's viewpoint as the front, which direction is the man facing in the image? A. front B. right C. back D. left	<b>Question:</b> From the perspective of the woman in blue clothes, where is the woman in white clothes located? A. left B. right C. front D. front-right
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	
	<b>Answer:</b> A. front		<b>Answer:</b> C. right	

Figure 9: ViewSpatial-Bench Examples (Part3).