ORIENTSAM: MITIGATING CAMERA-CENTRIC SHORTCUT IN MULTIMODAL SPATIAL REASONING VIA ORIENTATION-AWARE SPATIAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) still struggle with spatial reasoning that requires perspective transformation. In particular, they often rely on camera-centric cues rather than reasoning from the reference object's viewpoint, leading to systematic errors in non-camera reference settings. In this paper, we first analyze this failure mode and show that object orientation is a key factor underlying such camera-centric shortcut behavior. To address this issue, we propose **OrientSAM**, an orientation-aware spatial alignment framework for multimodal models. OrientSAM injects explicit orientation information into multimodal representations through orientation-aware tokens and Fourier-based angle encoding, and further adopts a curriculum learning strategy to progressively improve perspective-aware reasoning. In addition, we build a spatial data construction pipeline to generate orientation-aware spatial supervision from large-scale images. Experiments on Spatial-MM, ViewSpatial, and 3DSRBench show that OrientSAM delivers its strongest gains on non-camera-view, person-centric, and orientation-sensitive tasks. The results further demonstrate that explicit orientation modeling is important for mitigating camera-centric shortcut behavior and enabling more robust allocentric spatial reasoning in multimodal models.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved strong performance on vision-language tasks Liu et al. (2025); Yu et al. (2025), yet remain weak at spatial reasoning that requires perspective transformation. Many real-world tasks require relations to be judged from a reference object's local viewpoint rather than the image plane; for example, left/right judgments depend on both object positions and the reference object's facing direction Linsley et al. (2025); Li et al. (2025b); Liu et al. (2023); Rajabi & Kosecka (2024). This issue is central to embodied interaction Song et al. (2025), human-object collaboration Doğan et al. (2020), and scene understanding Wen et al. (2024); Bu et al. (2025), but current MLLMs are largely trained on camera-view image-text data and therefore tend to inherit camera-centric spatial bias.

Figure 1 shows a representative failure. Although the scene is clear, strong MLLMs follow the image-plane left-right layout instead of the reference player's local frame. Correctly answering requires first identifying the reference object's orientation and then transforming the target relation into that frame, suggesting that current models often use a camera-centered heuristic rather than perspective-aware reasoning.

We study this problem from both diagnostic and modeling perspectives. By disentangling the *reference-centric answer* from the image-plane *camera-centric answer*, we reveal a clear shortcut under conflict. Its strength changes with the reference object's orientation, indicating that missing orientation representation is a key bottleneck for reference-frame transformation.

Motivated by this finding, we propose **OrientSAM**, an **Orient**ation-aware **S**patial **A**lignment **M**odel. OrientSAM injects object orientation through orientation-aware tokens and Fourier angle encoding, supported by a multi-expert pipeline that combines region grounding, monocular geometry estimation, and orientation prediction. A curriculum then guides the model from orientation perception to perspective-aware spatial reasoning.

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

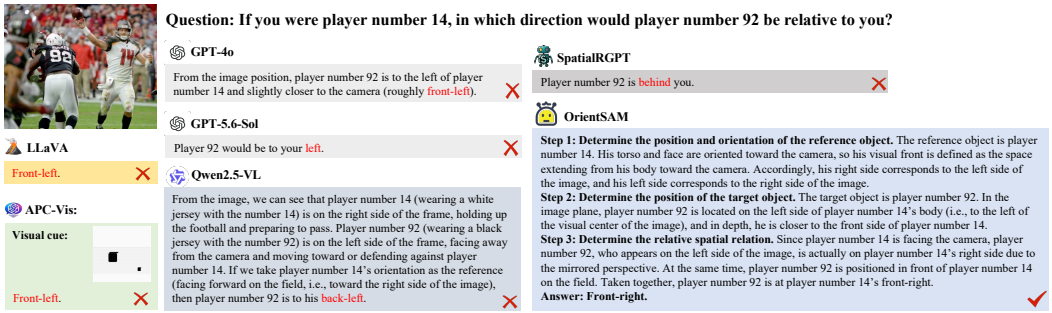


Figure 1: A representative case study. Existing models are biased toward camera-plane cues and fail to reason under the reference object's local viewpoint, whereas OrientSAM correctly predicts the spatial relation by incorporating the reference object's orientation.

Experiments on Spatial-MM, ViewSpatial, and 3DSRBench show that OrientSAM achieves leading results on the target non-camera-view, person-relative-direction, and orientation/direction tasks while reducing camera-centric shortcut behavior.

Our contributions are summarized as follows:

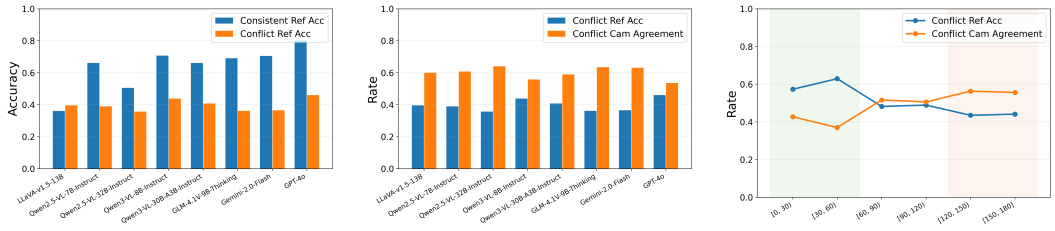
- We introduce a unified diagnostic framework that separates reference- and camera-centric answers, revealing a systematic camera-centric shortcut when the two answers conflict.
- We identify reference-object orientation as a key factor governing shortcut strength and motivate explicit orientation modeling for reference-frame transformation.
- We propose OrientSAM, combining multi-expert data construction, Fourier-encoded orientation tokens, multimodal alignment, and a two-stage curriculum for reference-centric reasoning.
- We achieve leading results on non-camera-view and orientation-sensitive subsets across three benchmarks, supported by ablations, counterfactual controls, and robustness analyses.

2 Related Work

Multimodal Spatial Reasoning. Spatial reasoning remains challenging for MLLMs when tasks require fine-grained grounding, metric depth, or 3D relations. Existing work improves these capabilities through large-scale spatial supervision and explicit geometric cues Cheng et al. (2025); Chen et al. (2025; 2024); Stogiannidis et al. (2025); Ogezi & Shi (2025); Wang et al. (2025b). SpatialRGPT Cheng et al. (2024) builds monocular 3D scene graphs with region and depth supervision, while SR-3D Cheng et al. (2025) unifies single- and multi-view reasoning using region prompts and canonical 3D positional embeddings. These methods target general spatial or 3D understanding, leaving orientation-conditioned reference-frame transformation underexplored.

Perspective-taking and Egocentric Bias. Prior analyses Lee et al. (2025); Leonard & Murray (2026); Chen et al. (2024); Kamath et al. (2023); Góral et al. (2024); Mayer et al. (2025); Yang et al. (2025) show that VLMs often default to egocentric or image-plane heuristics instead of allocentric reasoning. APC Lee et al. (2025) uses detection, segmentation, depth, and orientation to convert allocentric questions into egocentric ones, while token-based approaches Leonard & Murray (2026) inject embodiment and rotation cues directly into the model. We instead diagnose the behavior by separating camera-centric and reference-centric answers under conflict, exposing shortcut use that standard accuracy can hide.

3D-aware Representation and Orientation Modeling. Recent work increasingly emphasizes explicit 3D and orientation cues Huang et al. (2025); Lee et al. (2025); Jung et al. (2025): 3DRS Huang et al. (2025) distills pretrained 3D features, VGGT Wang et al. (2025a) predicts cameras and dense geometry, and Orient Anything Wang et al. (2025c) estimates semantic object orientation from single images. Together, they show the value of explicit geometry, but do not directly align orientation with reference-frame reasoning. OrientSAM targets this missing connection.



(a) Conflict Difficulty (b) Camera-Centric Bias (c) Orientation Dependence

Figure 2: Camera-centric shortcut analysis on ViewSpatial. (A) Conflict samples are harder than consistent ones. (B) Camera agreement exceeds reference accuracy under conflict. (C) The gap varies with orientation; rates are macro-averaged over the eight VLMs in (A–B).

Model	Consistent Ref Acc	Conflict Ref Acc	Conflict Cam Agree	Δ_{norm}
LLaVA-v1.5-13B	0.362	0.398	0.602	+0.513
Qwen2.5-VL-7B-Instruct	0.663	0.391	0.609	+0.557
Qwen2.5-VL-32B-Instruct	0.507	0.359	0.641	+0.786
Qwen3-VL-8B-Instruct	0.709	0.440	0.560	+0.273
Qwen3-VL-30B-A3B-Instruct	0.662	0.409	0.591	+0.445
GLM-4.1V-9B-Thinking	0.692	0.364	0.636	+0.747
Gemini-2.0-Flash	0.707	0.367	0.633	+0.725
GPT-4o	0.803	0.462	0.538	+0.165
Gemini-3.1-Pro Preview	0.791	0.472	0.528	+0.120
GPT-5.6 Sol	0.884	0.488	0.512	+0.048
Mean	0.678	0.415	0.585	**+0.438**

Table 1: Model-wise shortcut analysis on ViewSpatial. Positive Δ_{norm} indicates camera-centric bias. Δ_{norm} uses unrounded rates; the Mean row averages model-wise values.

3 Camera-Centric Shortcut in VLMs

We analyze whether MLLMs solve perspective-taking questions by reasoning in the reference object's local frame or by following image-plane cues. This diagnostic is necessary because the two strategies often agree on easy examples, making shortcut behavior invisible under accuracy alone.

Problem Setup. For each example, we define two candidate answers. The *reference-centric answer*, Ans_{ref}, is the ground-truth relation under the reference object's viewpoint. The *camera-centric answer*, Ans_{cam}, is the relation implied by the horizontal image-plane layout. We split examples into a *consistent* subset, where $\text{Ans}_{\text{ref}} = \text{Ans}_{\text{cam}}$, and a *conflict* subset, where $\text{Ans}_{\text{ref}} \neq \text{Ans}_{\text{cam}}$. The conflict subset directly tests whether a model performs reference-frame transformation rather than choosing the camera-view relation.

Metrics. Accuracy is always measured against Ans_{ref}. We additionally report *camera-centric agreement*, which measures whether the prediction matches Ans_{cam}. On conflict samples, high camera-centric agreement together with low reference-centric accuracy indicates reliance on the camera-centric shortcut. We summarize this tendency with

$$\Delta_{\text{norm}} = \frac{\text{CamAgree} - \text{RefAcc}}{\text{RefAcc}} \tag{1}$$

where larger positive values indicate stronger camera-centric bias.

Conflict samples are generally harder. Figure 2A shows that conflict samples are substantially harder than consistent ones for most models. Their errors therefore concentrate when image-plane cues diverge from the required reference-frame transformation, rather than arising uniformly from object localization.

Models prefer camera-centric answers under conflict. Figure 2B visualizes the original eight models, while Table 1 extends the comparison to ten models. All ten agree more often with Ans_{cam} than with Ans_{ref} on conflict samples. Across all ten models, mean reference-centric accuracy is 0.415, whereas camera-centric agreement reaches 0.585, and the mean model-wise Δ_{norm} is $+0.438$.

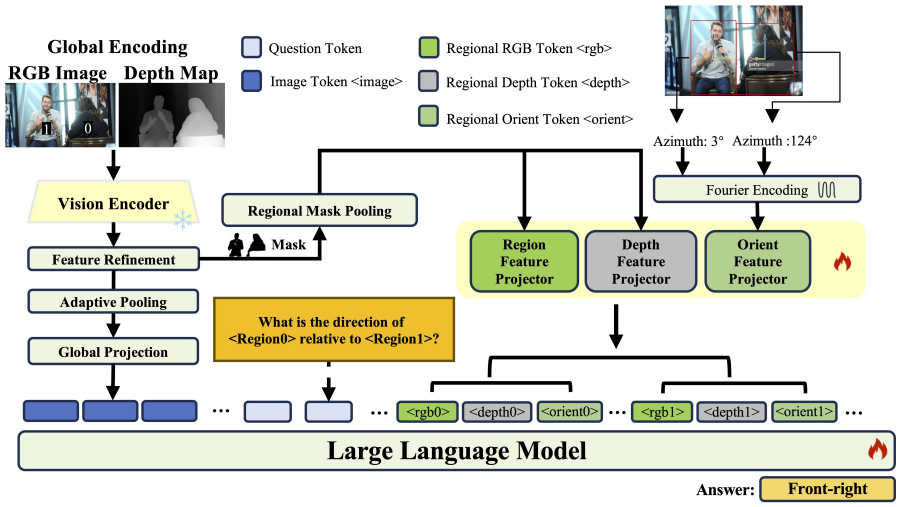


Figure 3: Overview of OrientSAM. The model augments multimodal reasoning with region-level RGB, depth, and orientation tokens. Object orientation is encoded with Fourier features and injected into the language model to support reference-centric spatial reasoning.

This consistent pattern indicates that the camera-centric shortcut remains widespread across both open-source and API models.

Shortcut strength depends on reference orientation. Figure 2C further shows that the gap varies across reference-orientation intervals. For each interval, we compute reference accuracy and camera agreement separately for each of the original eight VLMs and then macro-average the model-wise rates. Although not strictly monotonic, this orientation-dependent contrast links the shortcut to how models encode and use the reference object's facing direction during viewpoint transformation. Together, the results localize the failure to orientation-conditioned reference-frame transformation rather than generic spatial perception.

Importantly, aggregate accuracy can obscure this failure: a camera-centric prediction is still counted as correct whenever the two candidate answers coincide. Conflict examples remove this ambiguity, making the observed gap diagnostic of the required coordinate transformation rather than a uniform loss of visual recognition ability.

Implication for modeling. These findings motivate an explicit, learnable representation of reference-object orientation and its alignment with downstream spatial reasoning. We therefore propose OrientSAM, which injects orientation priors into multimodal representations and progressively learns perspective-taking through alignment and curriculum training.

4 Orientation-aware Spatial Alignment

OrientSAM integrates multi-expert processing, explicit orientation representation and alignment, and two-stage curriculum training for reference-centric spatial reasoning. As shown in Figure 3, the pipeline builds instance-aligned spatial descriptors, injects periodic orientation features alongside RGB and depth cues, and learns to use these cues through progressively harder supervision.

Problem Formulation. Given an image I, a query object o_q, and a reference object o_r, the task is to determine their spatial relation in the coordinate system of o_r. We extract object regions $\{o_i\}_{i=1}^{N}$ with bounding boxes and visual features, and associate each with an orientation representation ω_i encoding its semantic facing direction in 3D.

The task is to predict the relative spatial relation:

$$y = f(I, o_r, o_q, \omega_r) \tag{2}$$

where y is defined in the reference-centric coordinate system induced by ω_r, making the required transformation explicit rather than relying on image-plane coordinates.

Multi-expert Spatial Data Construction. We use a multi-expert pipeline both to construct offline training supervision and to produce the spatial inputs consumed at inference. At evaluation time, OrientSAM uses no ground-truth depth or orientation labels: following SpatialRGPT, the benchmark supplies the queried regions, while off-the-shelf estimators predict the corresponding depth and orientation cues before answer generation.

An instance extractor produces $O = \{o_i\}_{i=1}^{N}$, where $o_i = (b_i, m_i, t_i)$ contains the bounding box, segmentation mask, and semantic label.

The shared instance index keeps outputs from different experts attached to the same object, while the mask supports region-specific appearance and depth extraction for the queried reference frame.

A monocular geometry expert estimates scene structure and object depth as

$$(D, K, \xi) = F_{\text{geo}}(I), \qquad z_i = \text{Median}\{D(u, v) \mid (u, v) \in m_i\} \tag{3}$$

where D, K, and ξ denote depth, camera intrinsics, and camera pose, and z_i is the median mask depth.

An orientation expert predicts object pose and forms the unified representation

$$\begin{aligned}(\theta_i, \phi_i, \delta_i, c_i) &= F_{\text{orient}}(I, b_i) \\ \mathbf{v}_i &= \langle t_i, b_i, m_i, z_i, \theta_i, \phi_i, \delta_i, c_i \rangle\end{aligned} \tag{4}$$

where θ_i, ϕ_i, δ_i, and c_i are the polar angle, azimuth, rotation, and confidence.

We sample object pairs from these attributes to generate reference-centric questions and answers. Orientation-perception data retain predictions with $c_i > 0.98$; a manual audit of 200 non-camera-view samples found 93% correct grounding and reference-frame labels. Reasoning data comprise multi-object reference-centric pairs mixed with a moderate amount of general instruction data to mitigate catastrophic forgetting. These data support the two-stage curriculum described below.

Orientation-aware Representation. We explicitly encode object orientation in the multimodal representation to establish the reference object's local frame.

In the attribute vector $\mathbf{v}_i$ above, azimuth $\phi_i \in [0, 2\pi)$ characterizes the object's semantic facing direction.

Because a scalar angle is discontinuous at the $0/2\pi$ boundary, we preserve its periodic structure with Fourier features:

$$\gamma(\phi_i) = \left[\sin(2^k \phi_i), \cos(2^k \phi_i)\right]_{k=0}^{N-1} \tag{5}$$

where N is the frequency order.

Let $\mathbf{r}_i \in \mathbb{R}^{d_r}$ be the mask-pooled regional RGB feature and $\mathbf{d}_i \in \mathbb{R}^{d_d}$ the corresponding depth feature. We inject the orientation encoding into both the regional feature and its object token:

$$\tilde{\mathbf{r}}_i = \mathbf{r}_i + W_o \gamma(\phi_i), \qquad \tilde{\mathbf{e}}_i = \text{Embed}(t_i) + W_t \gamma(\phi_i) \tag{6}$$

where t_i is the semantic label or region identifier and W_o, W_t are learnable projections.

The multimodal encoder then combines the orientation-enhanced visual and textual features:

$$\mathcal{H} = \text{MMEncoder}\big(\{\tilde{\mathbf{r}}_i, \mathbf{d}_i\}_{i=1}^{N}, \{\tilde{\mathbf{e}}_i\}_{i=1}^{N}\big). \tag{7}$$

Thus, orientation becomes part of the multimodal context alongside appearance and depth, enabling reference-centric spatial reasoning.

Orientation-aware Alignment Learning. To make the geometric cues usable by the pretrained MLLM, we align RGB, depth, and orientation features within its language embedding space. Directly appending raw geometric values may leave the model dominated by pretrained RGB patterns; separate region-level tokens instead preserve each modality's role and make the reference object's facing direction explicitly accessible during reasoning.

Independent projectors map the three modalities of each object o_i into this space:

$$\mathbf{e}_i^{\text{rgb}} = W_r \tilde{\mathbf{r}}_i, \qquad \mathbf{e}_i^{\text{depth}} = W_d \mathbf{d}_i, \qquad \mathbf{e}_i^{\text{orient}} = W_\phi \gamma(\phi_i) \tag{8}$$

where W_r, W_d, and W_ϕ are learnable projection matrices.

Model	Spatial-MM		ViewSpatial			
	Cam	Non-Cam	C-Obj	C-Rel	P-Obj	P-Rel
Random Guess	37.61	39.93	25.00	25.00	25.00	30.14
LLaVA-v1.5-7B	41.34	40.30	26.50	27.24	24.00	27.91
LLaVA-v1.5-13B	51.24	46.27	25.60	28.14	23.39	28.15
Qwen2.5-VL-3B	64.66	32.09	30.22	40.67	38.55	30.76
Qwen2.5-VL-7B	77.21	37.31	26.00	44.61	37.65	32.66
Qwen2.5-VL-32B	78.80	38.81	31.53	57.66	56.85	39.07
Qwen3-VL-2B	77.56	35.82	23.80	48.11	44.98	38.84
Qwen3-VL-4B	82.51	32.09	31.22	49.63	43.07	40.50
Qwen3-VL-8B	82.69	38.81	31.83	53.13	51.57	39.55
SpatialRGPT	89.75	39.55	69.28	<u>78.62</u>	69.18	41.33
APC-Vis	<u>92.23</u>	<u>86.57</u>	<u>79.62</u>	73.32	**75.40**	72.57
Qwen3.7-Plus	88.63	47.67	46.37	65.33	62.48	52.09
Gemini-3.1-Pro Preview	90.17	53.61	51.48	69.47	66.23	57.56
GPT-5.6 Sol	**93.50**	67.02	59.31	77.13	<u>74.13</u>	<u>78.53</u>
OrientSAM	89.40	**87.31**	**83.33**	**80.49**	73.49	**85.04**

Table 2: Accuracy on Spatial-MM and ViewSpatial. Cam and Non-Cam denote camera- and non-camera-view; C and P denote camera- and person-view; Obj and Rel denote object orientation and relative direction. Bold and underline indicate the best and second-best result.

We preserve their distinct roles in structured region-level tokens and construct the final image-question sequence as

$$\begin{aligned} \mathcal{T}_i &= \{\langle rgb_i\rangle, \langle depth_i\rangle, \langle orient_i\rangle\} \\ \mathcal{X} &= \{\langle image\rangle, \mathcal{T}_1, \ldots, \mathcal{T}_N, \text{question tokens}\}. \end{aligned} \tag{9}$$

Here, $\langle rgb_i\rangle$ provides appearance, $\langle depth_i\rangle$ supports front–back disambiguation, and $\langle orient_i\rangle$ specifies the reference object's local coordinate system, allowing the model to attend to each modality as needed.

The model predicts the reference-centric answer y^{ref} with the standard next-token objective:

$$\mathcal{L}_{\text{NTP}} = -\log p_\Theta\left(y^{\text{ref}} \mid \mathcal{X}\right) \tag{10}$$

where Θ denotes all trainable parameters. We use no auxiliary loss against camera-centric answers; reference-centric supervision instead aligns the structured RGB–depth–orientation tokens with language reasoning.

Curriculum Learning Strategy. We separate orientation perception from reference-frame reasoning through curriculum training. Learning both simultaneously entangles facing-direction recognition with relation transformation and can favor an easier image-plane heuristic.

The two training objectives are

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}[\mathcal{L}_{\text{NTP}}^{\text{orient}}], \qquad \mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{NTP}}. \tag{11}$$

Stage 1: Orientation Perception. The model generates an azimuth answer for ϕ_i using next-token loss $\mathcal{L}_{\text{NTP}}^{\text{orient}}$, associating regional appearance with semantic facing direction.

Stage 2: Orientation-aware Reasoning. The model then uses the structured RGB–depth–orientation sequence to predict reference-centric relations. The curriculum thus establishes orientation sensitivity before full perspective-taking.

5 Experiments

Experimental Setup. We evaluate OrientSAM on **Spatial-MM** Shiri et al. (2024), **3DSRBench** Ma et al. (2024), and **ViewSpatial** Li et al. (2025a). Spatial-MM contains camera-centric and non-camera-centric spatial QA pairs; we use 566 camera-view and 134 non-camera-view pairs. 3DSR-Bench covers *Height* (690 samples), *Location* (1709), *Orientation* (1036), and *Multi-object Relation* (1722). ViewSpatial evaluates camera-view and person-view object orientation and relative direction. Because OrientSAM targets reference-frame transformation, our primary claims focus on

non-camera-view, person-centric, and orientation-sensitive subsets, while camera-view and height results serve as complementary diagnostics.

We compare with representative open-source MLLMs, including LLaVA-v1.5 Liu et al. (2024a), LLaVA-NeXT Liu et al. (2024b), Qwen2.5-VL Team (2025), and Qwen3-VL Bai et al. (2025); spatial reasoning models SpatialRGPT Cheng et al. (2024) and APC-Vis Lee et al. (2025); and recent API models Qwen3.7-Plus Qwen Team (2026), Gemini-3.1-Pro Preview Google (2026), and GPT-5.6 Sol OpenAI (2026). The 3DSRBench comparison additionally retains Gemini-2.0-Flash Google (2025) and GPT-4o OpenAI (2024). All tasks are evaluated by accuracy over the predefined answer.

5.1 Main Results

Results on Spatial-MM and ViewSpatial. Table 2 shows that OrientSAM is strongest where reference-frame transformation is required. On Spatial-MM non-camera-view questions, it reaches 87.31%, exceeding the second-best APC-Vis by 0.74 points. On camera-view questions, GPT-5.6 Sol and APC-Vis lead with 93.50% and 92.23%, while OrientSAM remains competitive at 89.40%, indicating that its non-camera-view gain does not substantially hurt camera-view reasoning.

On ViewSpatial, OrientSAM ranks first on three of four subtasks: camera-view object orientation, camera-view relative direction, and person-view relative direction. The largest gain is on person-view relative direction, where it exceeds the second-best GPT-5.6 Sol by 6.51 points (85.04% vs. 78.53%). On person-view object orientation, OrientSAM ranks third, 1.91 points below APC-Vis and 0.64 points below GPT-5.6 Sol. Its advantage is therefore concentrated on orientation-conditioned reasoning when camera-view cues are insufficient or misleading.

Model	Height	Loc.	Orient.	Multi.
Random Guess	50.00	50.00	37.56	41.69
LLaVA-v1.5-7B	50.14	64.37	42.17	45.16
LLaVA-v1.5-13B	57.25	61.26	39.16	42.66
LLaVA-NeXT-7B	54.20	57.87	39.45	42.83
Qwen2.5-VL-3B	51.59	58.86	41.59	43.87
Qwen2.5-VL-7B	57.97	65.12	50.77	52.59
Qwen2.5-VL-32B	52.17	62.84	41.12	41.87
Qwen3-VL-2B	50.94	66.94	43.62	44.26
Qwen3-VL-4B	52.39	73.87	48.15	50.02
Qwen3-VL-8B	54.78	73.93	49.09	50.22
SpatialRGPT	55.36	77.15	65.37	50.46
APC-Vis	54.49	56.24	<u>79.41</u>	47.04
Gemini-2.0-Flash	63.91	58.22	35.14	57.80
Gemini-3.1-Pro Preview	<u>77.00</u>	76.00	49.00	70.00
GPT-4o	74.35	67.00	35.24	58.86
GPT-5.6 Sol	**84.00**	<u>77.50</u>	62.00	<u>74.02</u>
OrientSAM	56.67	**78.23**	**81.83**	**77.80**

Table 3: Performance comparison on 3DSRBench. "Loc." denotes Location and Distance, "Orient." denotes Orientation and Direction, and "Multi." denotes Multi-object Relation.

Results on 3DSRBench. Table 3 confirms the same trend on 3DSRBench. OrientSAM ranks first on Location and Distance, Orientation and Direction, and Multi-object Relation, exceeding the corresponding second-best results by 0.73, 2.42, and 3.78 points. Height is a clear exception: GPT-5.6 Sol and Gemini-3.1-Pro Preview reach 84.00% and 77.00%, whereas OrientSAM obtains 56.67%. This contrast is consistent with the method's focus on reference-frame and orientation-aware reasoning rather than general height estimation.

5.2 Ablation Study

Effect of Orientation Feature Injection and Curriculum Learning. Table 4 reports the ablation results of orientation injection and curriculum learning. We compare four variants: Base-VLM (Zero-shot), Base-VLM (SFT), OrientSAM (w/o CL), and OrientSAM (Full). Base-VLM shares the Qwen2.5-VL-7B checkpoint with the OrientSAM backbone. Its zero-shot variant uses the region-conditioned input and prompt of the ablation pipeline, whereas the standalone Qwen2.5-VL-7B baseline in Table 2 follows the benchmarks' standard zero-shot interface. The results yield three observations. First, supervised fine-tuning alone brings only limited gains on reference-dependent tasks. Base-VLM (SFT) improves over the zero-shot baseline but remains weak on non-camera-view and person-centric reasoning, reaching 43.15% on Spatial-MM non-camera-view and 42.12% on ViewSpatial person-relative direction. Thus, data fine-tuning alone is insufficient for robust perspective transformation. Second, explicit orientation injection is the main source of performance gain.

Model	Comp.		Spatial-MM		ViewSpatial		3DSR
	Orient	CL	Cam	Non-Cam	P-Obj	P-Rel	Ori.-Dir.
Base-VLM (Zero-shot)	×	×	77.15	33.40	32.13	34.59	41.21
Base-VLM (SFT)	×	×	87.65	43.15	44.30	42.12	50.35
OrientSAM (w/o CL)	✓	×	88.54	76.29	65.40	74.32	72.83
OrientSAM (Full)	✓	✓	**89.40**	**87.31**	**73.49**	**85.04**	**81.83**

Table 4: Ablation results of orientation feature injection and curriculum learning (Accuracy, %). Comp. denotes components; Cam and Non-Cam denote camera-view and non-camera-view, respectively; P-Obj and P-Rel denote person-view object orientation and person-view relative direction, respectively; 3DSR denotes 3DSRBench, and Ori.-Dir. denotes orientation and direction.

Encoding	Spatial-MM		ViewSpatial		3DSRBench
	Camera-view	Non-camera-view	Person-Obj Ori.	Person-Rel Dir.	Ori. & Dir.
Scalar	86.53	75.26	62.41	72.13	73.52
Fourier ($N = 1$)	88.21	83.09	69.53	80.37	78.44
Fourier ($N = 2$)	**89.40**	**87.31**	**73.49**	**85.04**	**81.83**
Fourier ($N = 3$)	89.13	86.91	72.82	84.63	81.17

Table 5: Ablation study of different angle encoding strategies (Accuracy, %).

OrientSAM (w/o CL) substantially outperforms Base-VLM (SFT) across all orientation-sensitive tasks, including Spatial-MM non-camera-view (76.29% vs. 43.15%), ViewSpatial person-relative direction (74.32% vs. 42.12%), and 3DSRBench orientation & direction (72.83% vs. 50.35%). These improvements support the importance of explicit orientation-aware representation. Third, curriculum learning further improves performance. Compared with OrientSAM (w/o CL), the full model improves Spatial-MM non-camera-view from 76.29% to 87.31%, ViewSpatial person-relative direction from 74.32% to 85.04%, and 3DSRBench orientation & direction from 72.83% to 81.83%. These gains suggest that progressing from orientation perception to perspective-aware reasoning is more effective than optimizing all tasks in a single stage.

Effect of Angle Encoding Strategy. Table 5 compares different angle encoding strategies. Direct scalar encoding gives the weakest results, especially on non-camera-view and person-centric tasks, showing that naive scalar regression cannot properly model the periodic nature of orientation. Replacing scalar encoding with Fourier features consistently improves performance across all benchmarks. Among all variants, Fourier ($N = 2$) achieves the best overall results, including 87.31% on Spatial-MM non-camera-view, 85.04% on ViewSpatial person-relative direction, and 81.83% on 3DSRBench orientation & direction. Increasing the frequency order to $N = 3$ does not bring further gains, suggesting that second-order Fourier encoding provides the best balance between periodic continuity and discriminability.

Gain Attribution beyond Additional Data. Table 6 tests whether the gains come from extra instruction data or a simple text prompt. Same-data SFT reaches only 43.15%, and text-angle/text-frame prompting remains far below the full model. Shuffling the orientation token while fixing the image, question-answer pair, grounding, depth, training schedule, and token structure drops accuracy from 87.31% to 66.37%. This 20.94-point gap shows that correct instance-aligned orientation is necessary for the full improvement. The shuffled variant still exceeds same-data SFT by 23.22 percentage points, suggesting contributions from other structured spatial cues. Orientation therefore acts alongside these cues rather than explaining the entire improvement.

Setting	Non-Cam Acc.
Same-data SFT	43.15
SRGPT + text-angle prompt	41.23
SRGPT + text-frame prompt	46.38
OrientSAM with shuffled orient.	66.37
Full OrientSAM	**87.31**

Table 6: Counterfactual controls on Spatial-MM Non-Cam.

Robustness to Upstream Expert Errors. We perturb one test-time spatial input at a time while fixing the other cues, model weights, and evaluation samples. Region perturbations offset the box center and size by $\pm 5\%$, $\pm 10\%$, or $\pm 20\%$ of its scale. Depth estimates are shifted by ± 1 or ± 3 or-

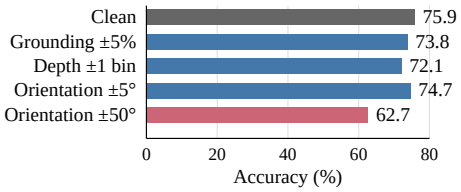


Figure 4: Robustness to upstream expert perturbations on 3DSRBench, measured by sample-weighted accuracy over all four task groups.

Model	Sec./sample	Acc.
Qwen2.5-VL-7B	**6.13**	37.31
SpatialRGPT	<u>10.21</u>	39.55
APC-Vis	18.13	<u>86.57</u>
OrientSAM	12.32	**87.31**

Table 7: End-to-end single-A100 inference on Spatial-MM Non-Cam.

dered bins, or have their ordering reversed. Orientation yaw is shifted by ± 1, ± 2, or ± 10 five-degree bins (i.e., $\pm 5°$, $\pm 10°$, or $\pm 50°$). The clean score of 75.9% is the sample-weighted overall accuracy across all 5,157 examples in the four 3DSRBench task groups. Figure 4 shows that mild region, depth, and orientation noise reduces this score to 73.8%, 72.1%, and 74.7%, respectively, while the severe orientation shift reaches 62.7%. The full sweep is reported in Appendix Table 10. This result shows robustness to errors in the test-time spatial inputs and motivates confidence filtering.

Inference Efficiency. Table 7 reports end-to-end single-A100 latency on Spatial-MM Non-Cam, including method-specific preprocessing, external experts, and answer generation. Despite its orientation estimation and spatial tokens, OrientSAM is faster and slightly more accurate than APC-Vis (12.32 vs. 18.13 seconds/sample; 87.31% vs. 86.57%).

5.3 Orientation-based Analysis

Figure 5 groups conflict samples by the reference object's coarse facing direction. For each dataset and facing group, we macro-average model-wise reference accuracy and camera agreement over the eight diagnostic VLMs. Front-facing samples exhibit higher camera agreement and lower reference accuracy than back-facing samples on Spatial-MM and ViewSpatial, with a smaller gap on 3DSR-Bench. This complements Section 3 by showing that shortcut strength varies with reference orientation rather than uniformly across spatial questions.

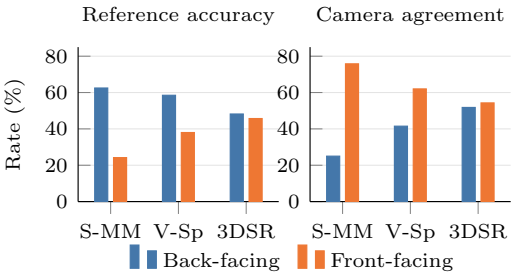


Figure 5: Orientation-conditioned conflict results macro-averaged over VLMs: reference accuracy (left) and camera agreement (right). S-MM: Spatial-MM; V-Sp: ViewSpatial.

Aggregate accuracy can conceal this pattern because camera-centric predictions remain correct when the two reference frames yield the same answer. Conflict samples and the facing-direction split expose orientation-dependent failures in reference-frame transformation. OrientSAM's gains on non-camera-view and person-view relative-direction subsets are therefore consistent with improved use of orientation, rather than a uniform gain on all visual questions.

6 Conclusion

We identify camera-centric shortcut behavior as an orientation-dependent failure mode in perspective-aware spatial reasoning, particularly when camera-centric and reference-centric answers conflict. OrientSAM addresses this issue by injecting Fourier-encoded orientation cues into region-level multimodal tokens and aligning them with spatial reasoning through curriculum learning. Across Spatial-MM, ViewSpatial, and 3DSRBench, its strongest gains occur on non-camera-view, person-centric, and orientation-sensitive tasks. These results support explicit orientation modeling for reducing camera-centric bias in multimodal spatial reasoning. The current formulation assumes a stable, visually identifiable front direction; extending it to ambiguous or dynamic reference frames remains an important direction.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing portions of the manuscript, language and presentation refinement, and literature retrieval and discovery. The authors checked retrieved references and reviewed all AI-assisted text for accuracy and consistency with the underlying work. The authors retain full responsibility for the final manuscript, including its technical content, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

The model architecture, orientation representation, training objectives, data-construction procedure, and experimental settings are described in the main paper. Additional definitions, implementation details, robustness analyses, and benchmark results are provided in the appendix to support reproduction and verification of the reported findings.

REFERENCES

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, Etash Kumar Guha, Silvio Savarese, Ludwig Schmidt, Yejin Choi, Caiming Xiong, and Ran Xu. BLIP3-KALE: Knowledge augmented large-scale dense captions. *CoRR*, abs/2411.07461, 2024. doi: 10.48550/arXiv.2411.07461.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yuqi Bu, Xin Wu, Zirui Zhao, Yi Cai, David Hsu, and Qiong Liu. Walk in others' shoes with a single glance: Human-centric visual grounding with top-view perspective transformation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26904–26923, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1306. URL https://aclanthology.org/2025.acl-long.1306/.

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

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *CoRR*, abs/2510.18632, 2025. doi: 10.48550/ARXIV.2510.18632. URL https://doi.org/10.48550/arXiv.2510.18632.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, Hongxu Yin, Xiaolong Wang, and Sifei Liu. 3d aware region prompted vision language model. *CoRR*, abs/2509.13317, 2025. doi: 10.48550/ARXIV.2509.13317. URL https://doi.org/10.48550/arXiv.2509.13317.

Fethiye Irmak Doğan, Sarah Gillet, Elizabeth J. Carter, and Iolanda Leite. The impact of adding perspective-taking to spatial referencing during human–robot interaction. *Robotics and Autonomous Systems*, 134:103654, 2020. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.

2020.103654. URL https://www.sciencedirect.com/science/article/pii/S0921889020304942.

Google. Gemini 2.0 flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.0-flash, 2025.

Google. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview, 2026. Accessed: 2026-07-24.

Gracjan Góral, Alicja Ziarko, Michal Nauman, and Maciej Wolczyk. Seeing through their eyes: Evaluating visual perspective taking in vision language models. *CoRR*, abs/2409.12969, 2024. doi: 10.48550/ARXIV.2409.12969. URL https://doi.org/10.48550/arXiv.2409.12969.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. In *Conference on Neural Information Processing Systems*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.03135.

Ji Hyeok Jung, Eun Tae Kim, Seo Yeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14257–14267. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01330. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jung_Is_Right_Right_Enhancing_Object_Orientation_Understanding_in_Multimodal_Large_CVPR_2025_paper.html.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.568. URL https://doi.org/10.18653/v1/2023.emnlp-main.568.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas J. Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *CoRR*, abs/2504.17207, 2025. doi: 10.48550/ARXIV.2504.17207. URL https://doi.org/10.48550/arXiv.2504.17207.

Bridget Leonard and Scott O. Murray. Cognitively-inspired tokens overcome egocentric bias in multimodal models. *CoRR*, abs/2601.16378, 2026. doi: 10.48550/ARXIV.2601.16378. URL https://doi.org/10.48550/arXiv.2601.16378.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqiao Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *CoRR*, abs/2505.21500, 2025a. doi: 10.48550/ARXIV.2505.21500. URL https://doi.org/10.48550/arXiv.2505.21500.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *CoRR*, abs/2510.12276, 2025b. doi: 10.48550/ARXIV.2510.12276. URL https://doi.org/10.48550/arXiv.2510.12276.

Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E. Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=UIFAJZ22ZF.

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Spatial intelligence in vision-language models: A comprehensive survey. *TechRxiv*, 2025(1110), 2025. doi: 10.36227/techrxiv.176231405.57942913/v2. URL https://www.techrxiv.org/doi/abs/10.36227/techrxiv.176231405.57942913/v2.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024c. doi: 10.1007/978-3-031-72970-6_3.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan L. Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *CoRR*, abs/2412.07825, 2024. doi: 10.48550/ARXIV.2412.07825. URL https://doi.org/10.48550/arXiv.2412.07825.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. iVISPAR — an interactive visual-spatial reasoning benchmark for VLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26757–26781, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1359. URL https://aclanthology.org/2025.emnlp-main.1359/.

Michael Ogezi and Freda Shi. SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7855–7875, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.387. URL https://aclanthology.org/2025.acl-long.387/.

OpenAI. Hello GPT-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. GPT-5.6 Sol Model. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026. Accessed: 2026-07-24.

Qwen Team. Qwen3.7: The agent frontier, May 2026. URL https://qwen.ai/blog?id=qwen3.7.

Navid Rajabi and Jana Kosecka. GSR-BENCH: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *CoRR*, abs/2406.13246, 2024. doi: 10.48550/ARXIV.2406.13246. URL https://doi.org/10.48550/arXiv.2406.13246.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 21440–21455. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1195. URL https://doi.org/10.18653/v1/2024.emnlp-main.1195.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 15768–15780. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01470. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_RoboSpatial_Teaching_Spatial_Understanding_to_2D_and_3D_Vision-Language_Models_CVPR_2025_paper.html.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *CoRR*, abs/2503.19707, 2025. doi: 10.48550/ARXIV.2503.19707. URL https://doi.org/10.48550/arXiv.2503.19707.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00499. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: towards spatial intelligence thorough evaluation. *CoRR*, abs/2505.05456, 2025b. doi: 10.48550/ARXIV.2505.05456. URL https://doi.org/10.48550/arXiv.2505.05456.

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

Chuan Wen, Dinesh Jayaraman, and Yang Gao. Can transformers capture spatial relations between objects? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=HgZUcwFhjr.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 10632–10643. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.23764.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are vlms from visual spatial intelligence? A benchmark-driven perspective. *CoRR*, abs/2509.18905, 2025. doi: 10.48550/ARXIV.2509.18905. URL https://doi.org/10.48550/arXiv.2509.18905.

A CAMERA-CENTRIC SHORTCUT ANALYSIS DETAILS

To further clarify the diagnostic framework used in Section 3, we provide additional details on how camera-centric shortcut behavior is defined and measured. Our analysis focuses on perspective-taking questions that ask for the spatial relation between two entities from the viewpoint of a reference object. The key question is whether a model truly performs reference-frame transformation, or instead follows a simpler camera-centric heuristic.

For each sample, we define two candidate answers. The first is the *reference-centric answer*, denoted by $\mathrm{Ans}_{\mathrm{ref}}$, which is the ground-truth spatial relation under the local viewpoint of the reference object. The second is the *camera-centric answer*, denoted by $\mathrm{Ans}_{\mathrm{cam}}$, which is the answer implied by the horizontal image-plane arrangement. In our setting, the answer space is restricted to binary left/right relations. Let x_q and x_r denote the horizontal center coordinates of the query object and the reference object, respectively. We compute $\mathrm{Ans}_{\mathrm{cam}}$ as

$$\mathrm{Ans}_{\mathrm{cam}} = \begin{cases} \texttt{left}, & x_q < x_r \\ \texttt{right}, & x_q > x_r. \end{cases} \tag{12}$$

This definition isolates the image-plane bias and allows us to compare it against the true reference-dependent answer.

Based on whether these two answers agree, we divide the samples into two subsets. A sample is called *consistent* if

$$\mathrm{Ans}_{\mathrm{cam}} = \mathrm{Ans}_{\mathrm{ref}} \tag{13}$$

and *conflict* otherwise:

$$\mathrm{Ans}_{\mathrm{cam}} \neq \mathrm{Ans}_{\mathrm{ref}}. \tag{14}$$

This split is central to the analysis. On consistent samples, the camera-centric answer and the reference-centric answer coincide, so a correct prediction does not necessarily indicate genuine perspective-taking. In contrast, conflict samples explicitly require the model to depart from the image-plane cue and reason under the reference object's local frame. Therefore, shortcut behavior becomes most visible on the conflict subset.

We evaluate model behavior from two complementary perspectives. The first is *reference-centric accuracy*, which measures whether the model prediction $\hat{y}$ matches the ground-truth answer:

$$\mathrm{RefAcc} = \mathbb{I}[\hat{y} = \mathrm{Ans}_{\mathrm{ref}}]. \tag{15}$$

The second is *camera-centric agreement*, which measures whether the prediction follows the image-plane relation:

$$\mathrm{CamAgree} = \mathbb{I}[\hat{y} = \mathrm{Ans}_{\mathrm{cam}}]. \tag{16}$$

Importantly, accuracy is always defined with respect to $\mathrm{Ans}_{\mathrm{ref}}$, while $\mathrm{Ans}_{\mathrm{cam}}$ is introduced only as a diagnostic variable. A model that truly performs reference-centric reasoning should achieve high RefAcc even when $\mathrm{Ans}_{\mathrm{cam}}$ disagrees with the ground truth. By contrast, if shortcut behavior exists, $\mathrm{CamAgree}$ will exceed RefAcc on the conflict subset.

On binary conflict samples with valid left/right predictions, reference accuracy and camera agreement satisfy $\mathrm{CamAgree} = 1 - \mathrm{RefAcc}$ when computed over the same examples. They therefore describe the same output behavior from opposite reference frames, rather than providing independent evidence. A direct model comparison requires a common conflict subset and scoring rule; benchmark-wide accuracy cannot be substituted for conflict-subset accuracy. These diagnostics characterize prediction behavior, not the model's internal reasoning mechanism.

For the model-wise analysis on ViewSpatial, we further report a normalized shortcut indicator:

$$\Delta_{\mathrm{norm}} = \frac{\mathrm{CamAgree} - \mathrm{RefAcc}}{\mathrm{RefAcc}} \tag{17}$$

computed on the conflict subset. A positive Δ_{norm} indicates that model predictions are more aligned with the camera-centric answer than with the reference-centric ground truth, and thus reflect a stronger shortcut tendency.

Following this framework, our shortcut analysis uses left/right person-perspective subsets drawn from Spatial-MM, ViewSpatial, and 3DSRBench. Each diagnostic sample is converted into a unified

Dataset	Model	Conflict RefAcc ↑	Conflict CamAgree ↓
Spatial-MM	Base-VLM (SFT)	0.40	0.60
	OrientSAM	0.80	0.20
ViewSpatial	Base-VLM (SFT)	0.42	0.58
	OrientSAM	0.76	0.24
3DSRBench	Base-VLM (SFT)	0.47	0.53
	OrientSAM	0.72	0.28

Table 8: Paired comparison on binary left/right conflict subsets.

format containing the image, the question, the reference-centric answer, the camera-centric answer derived from the image plane, the reference and query regions, and the estimated orientation of the reference object. This unified formulation allows the same diagnostic metrics to be applied across datasets without restricting the answer spaces used in the main benchmark evaluation.

For the orientation-grouped diagnostic plots, we compute reference accuracy and camera agreement separately for each of the original eight VLMs and then macro-average the model-wise rates within each angle or facing-direction group. This prevents models with larger conflict subsets from dominating the reported trend.

Beyond the consistent/conflict split, we also analyze how shortcut strength varies with the orientation of the reference object. Under the angle semantics adopted in our analysis, 0° corresponds to front-facing and 180° corresponds to back-facing relative to the camera. To measure the distance from the back-facing condition, we define

$$\theta_{\mathrm{abs}} = \mathrm{circ_dist}(\theta_{\mathrm{raw}}, 180^\circ) \tag{18}$$

where θ_{raw} denotes the original orientation angle and $\mathrm{circ_dist}(\cdot,\cdot)$ is the circular distance. We then divide the samples into six intervals:

$$[0, 30),\ [30, 60),\ [60, 90),\ [90, 120),\ [120, 150),\ [150, 180]. \tag{19}$$

In the paper, we additionally summarize the angle effect with two coarse regimes, namely back-facing $[0, 60)$ and front-facing $[120, 180]$, in order to study whether camera-centric shortcut behavior changes under different reference orientations.

This diagnostic framework supports three observations in the main text. First, conflict samples are generally harder than consistent ones, indicating that many models struggle when reference-centric reasoning diverges from image-plane cues. Second, among the original diagnostic baselines, camera-centric agreement is consistently higher than reference-centric accuracy on the conflict subset, which reveals a clear shortcut tendency. Third, the gap between these two quantities varies with the orientation of the reference object, suggesting that object orientation is an important factor underlying camera-centric shortcut behavior. In particular, the angle analysis is used as a diagnostic tool to study how shortcut strength changes across orientation conditions, rather than as an alternative definition of correctness.

A.1 Paired Comparison on Conflict Subsets

Table 8 compares Base-VLM (SFT) and OrientSAM on the same binary left/right conflict subset within each dataset. These diagnostic rates are distinct from the benchmark-wide accuracies in Tables 2 and 3. OrientSAM improves reference accuracy by 40, 34, and 25 percentage points on Spatial-MM, ViewSpatial, and 3DSRBench, respectively, with corresponding reductions in camera agreement. These results support reduced camera-centric answer preference on the evaluated conflict subsets; the two metrics are complementary rather than independent evidence.

B Multi-expert Spatial Data Construction

To alleviate the lack of supervision for perspective-aware spatial reasoning under non-camera reference frames, we build a multi-expert spatial data construction pipeline that automatically derives

spatial attributes and orientation-aware relations from large-scale images. The goal of this pipeline is to transform unstructured 2D image data into spatial question-answer pairs with explicit geometric grounding, so that the model can learn both object orientation perception and reference-frame transformation from scalable supervision.

The overall pipeline consists of five stages: image filtering, semantic tag extraction, 2D instance reconstruction, 3D geometric reconstruction, and automatic spatial annotation. Starting from large-scale image collections, we first retain only images that are suitable for constructing spatial reasoning samples. We then identify semantically meaningful entities in the image, localize them with region-level grounding, estimate their instance-level depth and orientation, and finally derive spatial relations under different reference frames. This design allows us to construct orientation-aware supervision without requiring manual 3D annotation.

In the first stage, we perform image filtering to select images with appropriate spatial structure. Since not all web images are suitable for perspective-aware reasoning, we retain only samples that satisfy several constraints, including a moderate number of salient entities, clear depth layering, recognizable semantic orientation, and relatively stable composition. In practice, we use an instruction-driven vision-language model to assess whether an image is appropriate for subsequent spatial annotation. This stage improves the overall quality of the generated data by removing images with severe ambiguity, excessive clutter, extreme camera tilt, or missing orientation cues.

In the second stage, for each retained image I, we extract a set of semantic tags

$$T = \{t_1, t_2, \ldots, t_M\} \tag{20}$$

where each t_i denotes a semantic category that appears in the image. These semantic tags serve as textual prompts for the following grounding stage, and provide the semantic anchor for constructing object-level spatial relations.

In the third stage, we reconstruct 2D instance-level information by combining text-guided detection and segmentation. Specifically, we use Grounding DINO Liu et al. (2024c) to obtain candidate bounding boxes conditioned on the extracted semantic tags, and then refine them with SAM Kirillov et al. (2023) to produce pixel-level masks. This yields an instance set

$$O = \{O_i\}_{i=1}^{n} \tag{21}$$

where each instance is represented as

$$O_i = (b_i, m_i, t_i). \tag{22}$$

Here, $b_i \in \mathbb{R}^4$ denotes the normalized 2D bounding box, m_i is the corresponding segmentation mask, and $t_i \in T$ is the semantic label. Since a single semantic tag may correspond to multiple object instances, the number of instances n is not necessarily equal to M.

In the fourth stage, we reconstruct 3D-related attributes from the monocular image. Given an input image I, the monocular geometry model predicts

$$(D_{\text{img}}, K, \xi) = F_{\text{geo}}(I) \tag{23}$$

where $D_{\text{img}} \in \mathbb{R}^{H \times W}$ is the dense depth map, $K \in \mathbb{R}^{3\times 3}$ is the estimated camera intrinsic matrix, and $\xi \in SE(3)$ denotes the camera pose. Based on the segmentation mask of each instance, we further compute an instance-level depth representation. To reduce the influence of local noise and outliers, we use the median depth inside the mask as the depth of instance O_i:

$$z_i = \text{Median}\left(\{D_{\text{img}}(u, v) \mid (u, v) \in m_i\}\right). \tag{24}$$

Besides depth, object orientation is another key attribute in our pipeline. For each instance region cropped from the image, we apply an orientation estimation model to obtain

$$(\hat{\theta}_i, \hat{\phi}_i, \hat{\delta}_i, c_i) = F_{\text{ori}}(\text{Crop}(I, b_i)) \tag{25}$$

where $\hat{\theta}_i$, $\hat{\phi}_i$, and $\hat{\delta}_i$ denote the predicted orientation parameters, and $c_i \in [0, 1]$ is the confidence score. The confidence score is used as a data quality signal to filter out samples whose semantic front direction is ambiguous or unreliable. In this way, the constructed data emphasize instances with stable and interpretable orientation semantics.

After the above steps, each entity is represented by a unified spatial attribute tuple:

$$V_i = \langle t_i, b_i, m_i, z_i, \hat{\theta}_i, \hat{\phi}_i, \hat{\delta}_i, c_i \rangle. \tag{26}$$

This representation combines semantic category, 2D localization, depth, and orientation, and serves as the basis for subsequent relation derivation.

In the final stage, we automatically generate orientation-aware spatial annotations. For a pair of entities, one is treated as the reference object and the other as the query object. Using their image-plane positions, instance-level depth, and the orientation of the reference object, we derive the relative relation under the reference-centered local frame. Compared with standard camera-view labeling, this step explicitly introduces reference-frame transformation into the annotation process. Therefore, the generated supervision is not limited to image-plane left/right relations, but instead reflects the spatial relation from the viewpoint of the reference object.

Based on the derived spatial attributes and relations, we construct spatial question-answer pairs for instruction tuning. These samples cover both basic orientation-related perception and more challenging perspective-aware reasoning under non-camera reference frames. In particular, the generated data are used together with the curriculum learning strategy described in the main text, so that the model can gradually learn from simpler orientation perception tasks to harder reference-frame transformation tasks.

The proposed multi-expert pipeline offers two main advantages. First, it avoids the need for expensive manual 3D annotation while still providing explicit supervision for orientation-sensitive spatial reasoning. Second, by integrating depth estimation, region grounding, and semantic orientation prediction into a unified process, it produces training data that are more directly aligned with the target reasoning behavior of our model. Although the quality of the generated annotations is still influenced by the accuracy of the expert models, this pipeline provides an effective and scalable way to construct orientation-aware spatial supervision from large-scale 2D images.

C Orientation-aware Representation and Training Details

A central goal of OrientSAM is to explicitly inject object orientation into the multimodal representation space, so that spatial reasoning can be performed under the local viewpoint of a reference object rather than the default camera plane. To this end, we augment the input representation with orientation-aware tokens and angle embeddings, and train the model with a progressive curriculum from basic orientation perception to perspective-aware spatial reasoning.

Given an image and a set of grounded regions, each object instance is associated not only with its semantic category and region information, but also with an estimated orientation. In our formulation, orientation is treated as a structured spatial attribute rather than a purely textual cue. Therefore, instead of only describing orientation in natural language, we explicitly encode it into the multimodal input sequence, allowing the model to access orientation priors at the representation level.

Concretely, for each target region, we introduce a dedicated orientation-aware placeholder token to indicate that the corresponding object is associated with orientation information. This token is used together with other region-related spatial markers in the input sequence, so that the model can distinguish ordinary semantic entities from entities carrying explicit geometric cues. In this way, orientation becomes part of the model's internal multimodal representation, rather than remaining an external textual hint.

To represent the orientation angle, we adopt Fourier-based angle encoding. A direct scalar encoding of angles suffers from discontinuity at the periodic boundary, where numerically distant values may correspond to geometrically similar directions. This issue is particularly undesirable for spatial reasoning, since nearby orientations should induce smooth changes in the representation space. To address this, we map the azimuth $\phi_i \in [0, 2\pi)$ into a continuous high-dimensional embedding using multi-frequency Fourier features:

$$\gamma(\phi_i) = \left[\sin(2^k \phi_i), \cos(2^k \phi_i)\right]_{k=0}^{N-1}. \tag{27}$$

Here, N denotes the frequency order, matching the notation used in the angle-encoding ablation. This mapping preserves the periodic structure of the angle variable and improves the smoothness and continuity of orientation-aware representations.

The resulting Fourier embedding is fused with the visual region representation, so that orientation information can participate in cross-modal alignment and downstream reasoning. Instead of treating angle as a separate symbolic label, this design enables the model to associate object appearance, spatial region, and orientation within a unified feature space. As a result, the model is encouraged to learn orientation-sensitive spatial patterns directly from representation interaction.

Based on this orientation-aware representation, the training process is further organized with a curriculum learning strategy. The motivation is that perspective-aware reasoning is a compositional ability: the model must first perceive the orientation of the reference object, and then use this information to transform the spatial relation into the corresponding local frame. Directly training on the final reasoning task alone makes this learning process unstable. Therefore, we adopt a progressive training schedule that gradually increases task difficulty.

At the early stage, the model is trained on simpler orientation-related perception tasks, which mainly strengthen its sensitivity to semantic front direction and orientation-dependent object understanding. This stage helps the model establish an initial correspondence between visual appearance and orientation embedding. After this basic capability is formed, the training data are gradually shifted toward more complex spatial reasoning samples that require reference-frame transformation. In these tasks, the model must combine region grounding, depth-related geometry, and the orientation of the reference object to infer the correct relation under a non-camera viewpoint.

This progressive strategy reduces the optimization difficulty of learning perspective-aware reasoning from scratch. In particular, it encourages the model to acquire orientation perception first, and then reuse this ability as a geometric prior for higher-level spatial reasoning. Such a staged training process is especially important in our setting, because the target reasoning behavior depends on both low-level direction understanding and high-level reference-frame transformation.

Overall, the orientation-aware representation and training design in OrientSAM consists of two tightly coupled components. The first is an explicit orientation-aware representation that encodes angle information through dedicated tokens and Fourier-based embeddings. The second is a curriculum learning schedule that organizes training from basic orientation perception to complex perspective-aware reasoning. Together, these two components allow orientation priors to be integrated into the multimodal hidden space in a smooth and learnable manner, thereby improving the model's robustness on orientation-sensitive spatial reasoning tasks.

D Additional Experimental Settings

D.1 Benchmark Evaluation and Diagnostic Subsets

We distinguish benchmark performance from shortcut diagnostics. The comparisons in Tables 2 and 3 retain the original answer space of each evaluated task, including object orientation, relative direction, height, location/distance, and multi-object relations. Predictions are scored against the corresponding task labels; these evaluations are not reduced to a common binary left/right task.

The consistent/conflict split and orientation-conditioned shortcut analysis are restricted to left/right questions defined under a reference object's viewpoint. Only these diagnostic samples receive both a reference-centric answer and a camera-centric answer derived from horizontal image-plane geometry. We apply the framework in Section 3 to these subsets, rather than to every question in the benchmarks. Orientation grouping changes the diagnostic partition, not the ground-truth answer or the main benchmark scoring rule.

All compared models answer the benchmark questions without ground-truth depth or orientation labels at test time. Following SpatialRGPT, queried regions are supplied by the benchmark; OrientSAM obtains the depth and orientation cues used during answer generation from off-the-shelf estimators.

D.2 Implementation Details

OrientSAM uses Qwen2.5-VL-7B as its backbone and injects orientation through the learnable region-feature and object-token projections described in Section 4. The RGB, depth, and orientation branches are projected into the language embedding space through W_r, W_d, and W_ϕ. The

Component	Setting
Backbone	Qwen2.5-VL-7B
2D instance reconstruction	Grounding DINO Liu et al. (2024c) and SAM Kirillov et al. (2023)
Instance depth	Median estimated depth within the instance mask
Orientation encoding	Azimuth in radians, $\phi_i \in [0, 2\pi)$; two Fourier bands
Training objectives	Next-token loss in both stages; orientation answers in Stage 1 and reference-centric relations in Stage 2
Training data	50K orientation-perception, 100K non-camera-view reasoning, and 400K auxiliary spatial/VQA examples
Optimizer and schedule	AdamW; peak learning rate 2×10^{-5}; cosine decay; 3% warmup; zero weight decay
Training batch	4 per device; 3 A100 GPUs; gradient accumulation 2; effective batch size 24
Training duration and seed	One epoch; seed 42; one run per trainable setting
Annotation filtering	Orientation confidence $c_i > 0.98$
Test-time spatial inputs	Queried regions with estimated depth and orientation; no ground-truth depth or orientation labels
Efficiency measurement	End-to-end, single A100; preprocessing, external experts, and answer generation included

Table 9: Representation, data, and inference settings for OrientSAM.

default angle encoding uses two Fourier frequency bands, yielding a four-dimensional embedding before projection. Table 9 summarizes the representation, data, and inference settings.

Optimization and infrastructure. Following the SpatialRGPT instruction-tuning recipe Cheng et al. (2024), we use AdamW with a peak learning rate of 2×10^{-5}, cosine learning-rate decay, a warmup ratio of 0.03, and zero weight decay. Training runs for one epoch using Python 3.10 and PyTorch 2.3.0 on three NVIDIA A100 GPUs. The per-device batch size is 4, with two gradient-accumulation steps, giving an effective global batch size of $4\times3\times2 = 24$. We use bfloat16 precision, TF32 computation, gradient checkpointing, and a random seed of 42. Unless otherwise specified, each trainable experimental setting is run once. The two curriculum stages use the orientation-perception and reference-centric reasoning objectives defined in Section 4; they do not correspond to SpatialRGPT's connector-initialization, pretraining, and instruction-tuning stages.

Orientation-perception objective. Stage 1 uses the same autoregressive next-token objective as instruction tuning, with orientation answers rather than spatial-relation answers as targets. Let s_i^{orient} be the tokenized azimuth answer and $\mathcal{X}_i^{\text{orient}}$ the image, referenced region, and orientation-question context. The per-example loss is

$$\mathcal{L}_{\text{NTP}}^{\text{orient}} = -\sum_{t=1}^{|s_i^{\text{orient}}|} \log p_\Theta\big(s_{i,t}^{\text{orient}} \mid \mathcal{X}_i^{\text{orient}}, s_{i,<t}^{\text{orient}}\big). \tag{28}$$

Thus, Stage 1 does not introduce a separate continuous-angle regression loss. Stage 2 changes the instruction and target to reference-centric relation reasoning while retaining autoregressive answer supervision.

D.3 Ablation Definitions and Interpretation

The component ablations use the Qwen2.5-VL-7B backbone. Base-VLM (Zero-shot) receives no spatial fine-tuning, whereas Base-VLM (SFT) is fine-tuned on the spatial instruction data without the orientation branch. OrientSAM (w/o CL) enables orientation modeling but uses a single-stage schedule; OrientSAM (Full) uses the two-stage orientation-perception and spatial-reasoning curriculum. Thus, the latter comparison contrasts training schedules with orientation modeling enabled in both variants.

The counterfactual controls address different questions. The SRGPT text-angle and text-frame variants test textual prompting as an alternative to OrientSAM, but use a different backbone and are not single-component ablations of the same model. The shuffled-orientation control instead preserves the orientation-token structure while disrupting the association between an object and its orientation.

Protocol	Perturbation Strength	Accuracy
Clean pipeline	None	75.9
Grounding: box center/size shift	$\pm 5\%/\pm 10\%/\pm 20\%$ of box scale	73.8 / 72.2 / 67.4
Depth: ordered-bin shift	$\pm 1/\pm 3$ bins; inversion	72.1 / 67.3 / 62.6
Orientation: yaw-bin shift	$\pm 1/\pm 2/\pm 10$ in 5-degree bins	74.7 / 72.4 / 62.7

Table 10: Cascaded-error stress test on 3DSRBench. Each perturbation is applied to one upstream expert stage at a time, and accuracy is sample-weighted over all four task groups.

Setting	Height	Location/Distance	Orientation/Direction	Multi-object
OrientSAM	56.67	78.23	81.83	77.80
De-duplicated data + retrained curriculum	56.32	78.56	80.13	76.47

Table 11: Results after CLIP nearest-neighbor de-duplication and curriculum-stage retraining on 3DSRBench.

Its comparison with full OrientSAM tests the importance of instance-aligned orientation, while the same-data SFT comparison addresses whether additional spatial instruction data alone reproduce the gains. These comparisons should not be interpreted as an exact additive decomposition of the contributions of data, representation, and curriculum.

E Robustness to Upstream Expert Errors

At inference, OrientSAM consumes the queried regions together with depth and orientation cues produced by the spatial preprocessing pipeline. To measure how errors in these inputs affect downstream reasoning, we independently perturb one input at a time while keeping the trained model and evaluation benchmark fixed. Region boxes are shifted in center and size, depth estimates are shifted across ordered bins or inverted, and orientation estimates are shifted across 5-degree yaw bins. Table 10 reports sample-weighted overall accuracy across all 5,157 examples in the four 3DSRBench task groups.

Small perturbations cause moderate degradation, whereas severe corruptions lead to larger drops. Thus, OrientSAM tolerates moderate errors in its test-time spatial inputs but inherits a failure boundary from the upstream estimators, motivating confidence-aware handling of uncertain predictions.

F Data Independence and Gain Attribution

The training images are drawn from general large-scale image collections, including BLIP3-KALE Awadalla et al. (2024) and LAION-5B Schuhmann et al. (2022), rather than from SpatialMM, ViewSpatial, or 3DSRBench images and test question-answer pairs. The 550K-example training set contains 150K constructed samples—50K orientation-perception samples and 100K non-camera-view reasoning samples—together with 400K auxiliary spatial/VQA examples. We retain orientation annotations with confidence $c_i > 0.98$. A manual audit of 200 non-camera-view examples found 93% to have correct reference/target grounding and consistent reference-frame question-answer labels.

We additionally apply source and image-ID filtering followed by CLIP nearest-neighbor de-duplication. We then retrain the curriculum stage using the filtered data. As shown in Table 11, performance remains close to the original result across all 3DSRBench subsets, suggesting that the observed gains are unlikely to be explained by near-duplicate overlap.

To examine gain attribution, we compare textual orientation prompting with an instance-alignment control. The SRGPT prompting variants verbalize either the orientation angle or the induced front/back/left/right frame. The shuffled-orientation variant keeps the image, question-answer pair, grounding, depth, training schedule, and token structure unchanged, but replaces each orientation token with one sampled from another example in the same object-category group. The prompting

Model	OmniSpatial			MMSI
	Ego	Allo	Hypo	
Qwen2.5-VL-7B	64.5	33.2	37.4	25.9
OrientSAM	67.1	44.6	43.9	27.0

Table 12: Additional spatial benchmark results.

alternatives and the within-model shuffle address distinct comparisons, as detailed in the ablation definitions above.

Table 6 shows that textual orientation prompts alone do not reproduce the full improvement. Shuffling the orientation signal also reduces accuracy by 20.94 points relative to full OrientSAM. This gap supports the contribution of correct, instance-aligned orientation information, while the remaining improvement over same-data SFT suggests that other structured spatial cues also contribute.

G Efficiency and External Benchmark Evaluation

Table 7 reports end-to-end single-A100 latency on Spatial-MM non-camera-view examples, including method-specific preprocessing, external experts, and answer generation. OrientSAM adds computation over Qwen2.5-VL-7B and SpatialRGPT but remains faster and slightly more accurate than APC-Vis.

We also test transfer on OmniSpatial's three perspective-taking splits Jia et al. (2026) and MMSI-Bench Yang et al. (2026). Table 12 shows the clearest gain on OmniSpatial allocentric reasoning, from 33.2 to 44.6, with smaller gains elsewhere.

Together, these results calibrate the method's scope: OrientSAM targets reference-frame transformation, with its strongest gains on direction-sensitive, non-camera-view, and allocentric reasoning rather than generic visual perception.

H Limitations and Discussion

Although OrientSAM improves perspective-aware spatial reasoning by explicitly incorporating orientation information, several limitations remain. First, the framework relies on external spatial estimators during both data construction and inference, including monocular geometry and object orientation prediction. Their errors can affect the automatically generated supervision and propagate through the test-time spatial inputs to the final answer.

Second, the effectiveness of the proposed approach depends on whether the reference object has a stable and semantically meaningful front direction. In real images, some objects have ambiguous orientation, weak visual cues, or heavy occlusion, making their semantic facing direction difficult to estimate reliably. In such cases, even explicit orientation modeling may provide limited benefit, since the reference frame itself is not well defined.

Third, the shortcut diagnostic focuses on binary left/right relations under a reference object's viewpoint. Although the main benchmark evaluation includes height, location/distance, orientation, and multi-object tasks, these broader accuracy results do not by themselves establish reference-frame transformation for every relation type. Reasoning under ambiguous or dynamic local frames remains outside the scope of the current diagnostic study.

Despite these limitations, our results suggest that object orientation is an important missing factor in current multimodal spatial reasoning systems. The analysis in the main paper shows that existing models often default to camera-plane heuristics when reference-centric reasoning conflicts with image-plane cues, and that this shortcut behavior is closely related to the orientation of the reference object. From this perspective, our method should be viewed not only as a performance-oriented improvement, but also as an attempt to introduce a more appropriate spatial inductive bias into multimodal representations.

More broadly, the proposed framework highlights the importance of explicitly modeling local reference frames in multimodal reasoning. Many practical scenarios, such as human-object interaction and scene understanding, require the model to reason from the viewpoint of an entity in the scene rather than from the camera alone. Our findings indicate that enhancing multimodal models with orientation-aware representations and progressively trained perspective transformation ability is a promising step toward more robust allocentric spatial reasoning.

In future work, it would be valuable to reduce dependence on external spatial estimators, extend the formulation to richer spatial relation types, and study how orientation-aware reasoning can be integrated into broader embodied or interactive multimodal settings.