Lost in Space? Vision-Language Models Struggle with Relative Camera Pose Estimation

Anonymous ACL submission

Abstract

We study whether vision-language models (VLMs) can solve relative camera pose estimation (RCPE) from image pairs, a direct test of multi-view spatial reasoning. We cast RCPE as a discrete verbal classification task and introduce `VRRPI-Bench`, built from real RGB-D frames with object-centric camera motion, and `VRRPI-Diag`, which isolates individual motion degrees of freedom. Humans (0.91) and specialized geometric pipelines such as LoFTR (0.99) solve the task reliably, yet the best VLM reaches only 0.66 and most others remain near random. Our analyses show that this gap is not basic spatial competence: strong VLMs are near ceiling on single-image benchmarks, but most remain near random once reasoning must span views. They are unstable under source-target reversal (best 59.7% consistency) and remain weak even in simplified single-DoF settings, especially on optical-axis motions such as roll and depth translation (GPT-5: 0.46 on roll). These failures are useful: they localize concrete missing capabilities, namely cross-view correspondence, view-consistent reasoning, and projective camera-motion understanding, making RCPE a targeted diagnostic for improving multi-view spatial reasoning in VLMs. [1]

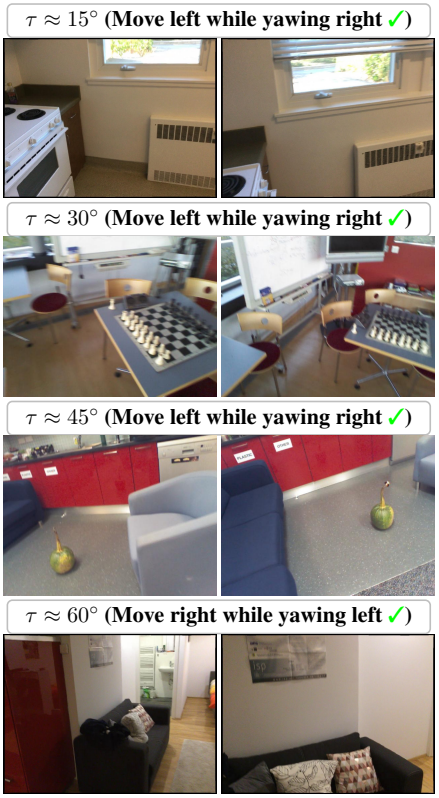

Figure 1: Examples from `VRRPI-Bench`. Camera moves from source observation (left) to target observation (right), with the verbalized camera motion for simultaneous translation and rotation around a shared object.

1 Introduction

Relative camera pose estimation (RCPE), which aims to recover the relative rotation and translation between two camera viewpoints, is a fundamental problem in computer vision and robotics (Kelsey et al., 2006; Zhou and Roumeliotis, 2008; Elqursh and Elgammal, 2011). While a single image provides only a projective and inherently ambiguous view of the physical world, multiple images enable reliable 3D perception once their relative pose is known. Beyond its classical formulation, RCPE represents the minimal unit of motion estimation from visual observations, serving as a structural building block for higher-level perception systems, including navigation (Cassinis et al., 2019), augmented reality (Marchand et al., 2015), and 3D reconstruction (Mouragnon et al., 2006; Hoang et al., 2020), accurate estimation of relative camera motion is a prerequisite for system initialization, tracking, and re-localization.

Traditional RCPE approaches rely on geometric constraints from low-level correspondences, such as keypoint matching and epipolar geometry (Lowe, 2004; Sun et al., 2021; Rao et al., 2026). While

[1] The code and benchmarks will be publicly available.

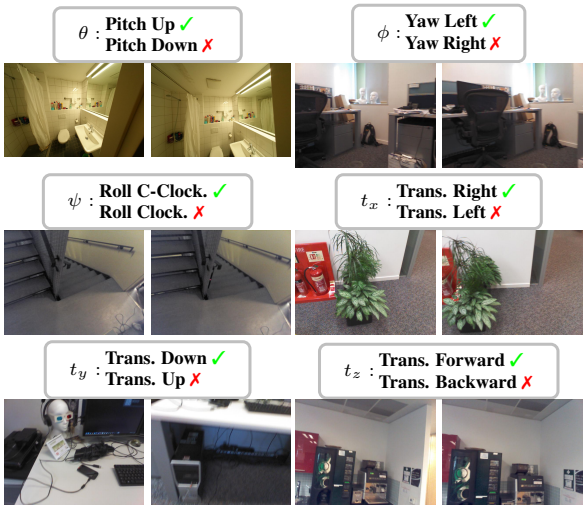

Figure 2: Examples from VRRPI-Diag. Camera moves from source viewpoint (left) to target viewpoint (right) and both positive and negative descriptions are provided.

effective under moderate viewpoint changes, they often degrade under wide baselines, occlusions, or texture-poor regions. Humans remain relatively robust, suggesting that higher-level semantic and structural cues help resolve spatial ambiguities. This motivates asking whether models with strong semantic priors can complement or extend classical geometry-based pipelines.

In parallel, VLMs have rapidly advanced and are increasingly deployed in embodied and interactive systems, particularly in mobile robotics (Shah et al., 2023). Such systems operate over time and across changing viewpoints, requiring perceptual representations that remain consistent under camera motion. This naturally calls for reasoning about spatial relationships across multiple views rather than within a single static image. Recent studies have begun to evaluate spatial reasoning, but primarily focus on single-image 2D relations, synthetic scenes, or highly constrained 3D settings (Kamath et al., 2023; Chen et al., 2024). These benchmarks do not directly assess whether VLMs can reason about metric, multi-view geometry grounded in real images. As a result, modern VLMs' ability to infer camera motion remains unexplored.

We investigate whether state-of-the-art VLMs can infer relative 3D camera pose from pairs of real-world images. To facilitate this evaluation, we introduce VRRPI-Bench (**V**isual **R**easoning **R**elative **P**ose **I**dentification), a benchmark constructed from unlabeled egocentric videos with annotations describing verbalized changes in relative camera translation and rotation (§3). The benchmark reflects realistic scenarios in which two views predominantly focus on the same object while undergoing simultaneous translation and rotation (see Figure 1). To make RCPE compatible with current multimodal models, we formulate the task as a discrete classification problem that captures the dominant direction of camera motion rather than predicting a full continuous transformation. This formulation preserves the geometric essence of RCPE while enabling systematic evaluation of VLMs' 3D and multi-view spatial reasoning abilities.

Our experiments show existing VLMs (best 0.66) lag behind both humans (0.91) and classical geometric methods (0.99) on RCPE (§5.1). We further demonstrate that VLMs are inconsistent (best 59.7%) when reasoning over image pairs, failing to preserve logical symmetry when source and target views are swapped (§5.2). To understand the source of this gap, we introduce a diagnostic benchmark, VRRPI-Diag, which isolates single degree of freedom camera motions for fine-grained analysis (see Figure 2). Using this benchmark, we show that VLMs struggle most with transformations along the optical axis (GPT-5 achieves 0.90 on average, but 0.46 on roll prediction), particularly depth translation and roll rotation (§5.3).

Finally, we conduct an error analysis to disentangle perception, cross-view correspondence, and camera-perspective reasoning (§6). We show that while modern VLMs have largely solved basic single-image spatial relations (most VLMs over 90, compared to 46.8 by BLIP2), they fail to robustly track objects across viewpoints (most VLMs near random level 0.28) and to internalize the inverse relationship between object motion and camera motion. Together, these findings indicate that current VLMs rely on shallow, image-plane heuristics rather than grounded 3D geometric reasoning. We argue that RCPE, and VRRPI-Bench in particular, provides a stringent and complementary testbed for evaluating spatial intelligence in VLMs. Progress on this task would signal meaningful advances in multi-view understanding and grounded 3D reasoning, which are prerequisites for deploying VLMs in real-world embodied systems.

2 Related Work

Pose Estimation RCPE is a fundamental problem in computer vision, especially simultaneous localization and mapping (SLAM) (Kelsey et al.,

2006; Zhou and Roumeliotis, 2008; Elqursh and Elgammal, 2011). Classical methods are divided into two categories: (i) **Solver-based**: SIFT (Lowe, 2004), SURF (Bay et al., 2006), and LoFTR (Sun et al., 2021) are used to identify matching points across views, where a matrix solver (Fischler and Bolles, 1981) computes the relative pose, emphasizing precision; (ii) **Learning-based**: Rockwell et al. (2022); Khatib et al. (2024) use deep networks to estimate pose directly, focusing on robustness. FAR (Rockwell et al., 2024) presents a hybrid approach, achieving state-of-the-art performance.

Enhancing Spatial Reasoning of VLM The common trends in enhancing spatial reasoning capabilities can be summarized through several approaches: (i) **Multi-modal Input**: Spatial-RGBT (Cheng et al., 2024) and SSR (Liu et al., 2025a) incorporate additional modules to handle extra-modal data, such as depth images and point clouds, to enhance 3D space understanding; (ii) **Augmented-VLM**: SpatialPIN (Ma et al., 2024) and MM-Spatial (Daxberger et al., 2025) introduce external tools to aid in constructing 3D scenes, while Cola (Chen et al., 2023) augments VLMs with an agent that acts as a coordinator; (iii) **Training**: VILASR (Wu et al., 2025b) uses both supervised fine-tuning and reinforcement learning, Spatial-MLLM (Wu et al., 2025a) modifies the vision-encoder architecture and releases their pre-trained VLM, and SpatialCoT (Liu et al., 2025b) uses supervised learning to align VLMs with coordinate representations for improved reasoning.

Spatial Reasoning Benchmark Spatial reasoning in VLM is evaluated across several dimensions: (i) **2D Reasoning**: Sparkle (Tang et al., 2025) and SpatialEval (Wang et al., 2024) build upon setups from LLM multi-hop reasoning tasks, treating image inputs as illustrations of grids, maps, or mazes; (ii) **3D Reasoning**: Chen et al. (2024); Daxberger et al. (2025); Song et al. (2025); Cai et al. (2025) focus on understanding 3D scenes, emphasizing coordinate alignment, metric estimation (Liao et al., 2024), and spatial relationships (Kamath et al., 2023; Du et al., 2024); (iii) **Mental Modelling**: Mind the Gap (Stogiannidis et al., 2025) and MindCube (Yin et al., 2025) introduce a challenging task involving mental inference (reasoning on unobservable processes based on limited views); (iv) **Comprehensive**: Fu et al. (2024); Shiri et al. (2024); Yang et al. (2025) propose benchmarks that integrate the above tasks. As

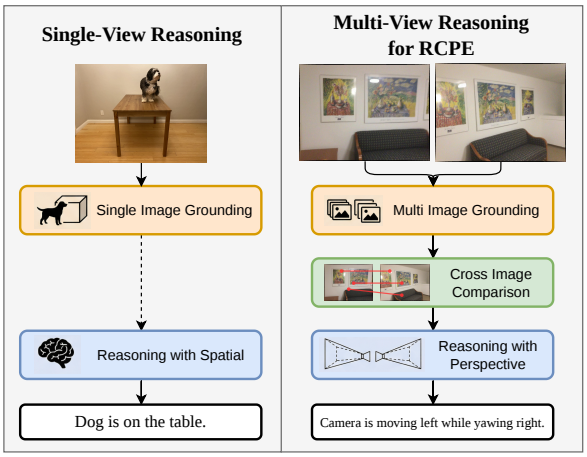

Figure 3: Capability taxonomy highlighting the skills required by VRRPI-Bench. Multi-view reasoning (right) raises new challenges in comparing with similar images, compared to single-view reasoning (left).

illustrated in Figure 3, unlike previous spatial reasoning tasks, our benchmark involves cross-image comparison, where models must reconcile spatial relations across viewpoints. VRRPI-Bench evaluates a richer set of spatial reasoning skills than prior benchmarks. See App. A.3 for more comparison.

3 Data Collection

3.1 Preliminaries

Relative Pose Vector We use **6 DoF** (degrees of freedom) to describe the camera's movement in a 3D scene: both rotation and translation along the x, y, and z axes. For rotation, we use pitch (θ) around the x-axis, yaw (ϕ) around the y-axis, and roll (ψ) around the z-axis. For translation, we represent movement along the x-, y-, and z-axes as t_x, t_y, and t_z, respectively. Therefore, we use $\boldsymbol{v} \triangleq (\theta, \phi, \psi, t_x, t_y, t_z)$ to represent the relative pose vector, which corresponds to the 6 DoF. The rotation is expressed in radians [2], while the translation follows the scale of the dataset metric units.

Intrinsic and Extrinsic Parameters The camera's *intrinsic matrix* is denoted as $\mathbf{K} \in \mathbb{R}^{3\times3}$ and contains the pinhole camera projection parameters. The *extrinsic matrix*, representing the camera's pose, is denoted as $\mathbf{T}_{wc}$, which is a homogeneous transformation matrix mapping points from the *camera coordinate system* $\{c\}$ to a universal *world coordinate system* $\{w\}$. Specifically,

[2] Although rotations are represented in radians internally, we report and discretize angular differences in degrees for evaluation.

a point $\mathbf{p}_c$ expressed in $\{c\}$ is transformed to $\{w\}$ by $\mathbf{p}_w = \mathbf{T}_{wc}\mathbf{p}_c$. The extrinsic matrix $\mathbf{T}_{wc}$ can be expressed as $\mathbf{T}_{wc} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3)$, where $\mathbf{R} \in SO(3)$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector.

3.2 Filtering Algorithm

Next, we describe two filtering schemes that we use in our experiments: `VRRPI-Bench` and `VRRPI-Diag`, representing global camera movement and single DoF change respectively. More technical details are provided in App. A.

Curation of `VRRPI-Bench` For i-th RGB-D frame $\mathcal{F}_i$, we are given an RGB image I_i, a depth map[3] D_i, the camera intrinsic matrix $\mathbf{K}_i$, the camera index c_i, and the camera-to-world pose $\mathbf{T}_{wc_i}$, i.e., $\mathcal{F}_i = (I_i, D_i, \mathbf{K}_i, \mathbf{T}_{wc_i})$, we denote the entire dataset as $\{\mathcal{F}_i\}_{i=1}^n$, where all frames share a universal world coordinate system w. The filtering algorithm selects image pairs observing the same central object from different viewpoints, ensuring geometric consistency. For each pair (I_i, I_j), we compute the angular difference τ and the mean reprojection distance $\bar{d}$ and keep it only if $\bar{d} < d^+$, where d^+ is a predefined maximal reprojection error threshold. We then define the **angular ranges** τ around a predefined angle set: $15°, 30°, 45°, 60°$ representing different task levels and classify the pair into one of the segments. Intuitively, by maintaining the same 3D central point in the middle of both images, we ensure meaningful object-centric camera rotation. Note that such a movement usually requires a two DoF change in the extrinsic camera matrix. This process produces viewpoint-consistent samples for `VRRPI-Bench`.

Curation of `VRRPI-Diag` For i-th frame $\mathcal{F}_i$, we consider only the RGB image I_i and its associated camera-to-world pose $\mathbf{T}_{wc_i}$, i.e., $\mathcal{F}_i = (I_i, \mathbf{T}_{wc_i})$. we denote the dataset as $\{\mathcal{F}_i\}_{i=1}^n$. Following the filtering principle of `VRRPI-Bench`, this algorithm selects image pairs that emphasize *single-DoF motion* (translation or rotation along one axis). For each pair (I_i, I_j), we compute the relative transformation $\mathbf{T}_{c_ic_j}$. A pair is retained if exactly one of the six DoF motions exceeds a threshold δ^+, while the remaining DoF motions stay below δ^- (predefined per-DoF thresholds). This process produces single-DoF-motion samples for `VRRPI-Diag`.

[3]In depth map, each pixel value represents the camera-to-scene distance in a camera-dependent metric unit.

	τ			
	~15°	~30°	~45°	~60°
# Frames	564	413	350	186

Table 1: Statistics for `VRRPI-Bench`.

3.3 Overview

The dataset sources that we use to collect data points are:

- **7 Scenes** (Shotton et al., 2013) is an indoor RGB-D dataset (~43k) from Microsoft Research with seven environments and precise camera poses.
- **ScanNet** (Dai et al., 2017) is a large-scale indoor dataset containing 2.5M frames across 1500 scenes with RGB-D information.
- **ScanNet++**[4] (Yeshwanth et al., 2023) contains 460 indoor scenes captured by laser scanners and commodity-level RGB-D cameras.

In summary, each benchmark sample consists of a pair of RGB images paired with a natural language description of the relative camera motion, sourced from RGB-D frames with known camera poses. Statistics about `VRRPI-Bench` are shown in Table 1, while the statistics of `VRRPI-Diag` can be found in App. A.3 (Table 5). After filtering, we obtain an average of ~190 samples per difficulty level across all datasets.

4 Experimental Setup

Task-Specific Baselines We use SIFT (Lowe, 2004) and LoFTR (Sun et al., 2021) as baseline methods for feature matching, followed by RANSAC (Fischler and Bolles, 1981) to estimate the essential matrix and compute the relative pose. SIFT identifies distinctive keypoints (unique locations within an image based on pixel values) that can be matched across different images, while LoFTR uses a vision transformer to directly match corresponding points (locations that appear in multiple images) between images. More details about evaluation can be found in App. B.1

Vision-Language Model We evaluate a range of VLMs and group them into four categories: 1) **Open-source**: Idefics3 (Laurençon et al., 2024), LLaVA-NeXT (Liu et al., 2024), LLaVA-OneVision (Li et al., 2024), Qwen2.5-VL (Bai

[4]Depth maps and intrinsics are unavailable for ScanNet++, so we only use it for `VRRPI-Diag`.

et al., 2025b), and Qwen3-VL (Bai et al., 2025a); 2) **Proprietary**: GPT-4o (OpenAI et al., 2024) and GPT-5 (Singh et al., 2025); 3) **Fine-tuned VLM**: SpaceQwen (Chen et al., 2024), based on Qwen2.5-VL and fine-tuned on SpatialVLM; and 4) **Reasoning VLM**: GLM-4.1V-Thinking (Team et al., 2025) and Qwen3-VL-thinking, designed for step-by-step reasoning. See App. B.2 for prompts we use and evaluation details.

Human Evaluation We randomly sample 40 examples from `VRRPI-Bench` and ask human annotators (25 graduate students) to follow a prompt matching the one provided to the VLMs, without any supplementary information or examples. Further details on the prompt and evaluation procedure are provided in App. B.3.

Evaluation We report **macro F1-score** to assess correctness and robustness of VLM performance.

5 Results

5.1 Main Results on `VRRPI-Bench`

LoFTR nearly saturates the 7 Scenes subset (0.99) and maintains high robustness on ScanNet (0.92), while SIFT remains a strong baseline, outperforming most of VLMs on both benchmarks (Table 2). Geometric methods show a pattern: performance drops with increasing viewpoint angles as local feature correspondences become harder. However, human annotators actually perform better at larger angles, as wider baselines require more global reasoning. VLMs, despite their semantic strengths, do not leverage global structural cues effectively to handle viewpoint shifts, often falling below random-level performance. This suggests that while humans use global priors to resolve spatial ambiguity, VLMs lack the geometric reasoning needed to exploit scene-level information. As a matter of fact, in several cases, where performance is low, the model returns degenerate answers, with a strong prior towards one answer that repeats and is also unaffected by the order of the images (§5.2).

Model scale provides only modest improvements, with even the strongest open models, such as the largest Qwen variants, falling well short of geometry methods. `GPT-5` stands as the clear exception, delivering the best VLM performance overall (0.66 on 7 Scenes and 0.64 on ScanNet) and demonstrating the most stable behavior as the angle increases. On ScanNet, its accuracy at large rotations (0.62 to 0.66) approaches that of traditional matchers, suggesting that VLMs show a potential in wide-baseline scenarios where geometric methods struggle. Despite this progress, a gap to human performance remains. Overall, the results emphasize both the current limitations of VLMs and the emerging potential of next-generation models in large-angle settings.

	7 Scenes					ScanNet				
Angle τ	15°	30°	45°	60°	Avg.	15°	30°	45°	60°	Avg.
SIFT	0.96	0.82	0.77	0.83	0.89	0.77	0.61	0.55	0.57	0.63
LoFTR	1.00	0.99	0.96	0.95	0.99	0.96	0.95	0.90	0.82	0.92
Lv-Next-8B-Lm3	0.40	0.43	0.42	0.38	0.41	0.42	0.40	0.40	0.38	0.40
Lv-OneVision-7B-Q2	0.35	0.35	0.35	0.34	0.35	0.37	0.34	0.36	0.30	0.35
Idefics3-8B-Lm3	0.47	0.48	0.43	0.40	0.47	0.49	0.51	0.52	0.53	0.51
SQ2.5-VL-3B-I	0.43	0.35	0.42	0.41	0.41	0.37	0.47	0.39	0.45	0.42
Q2.5-VL-3B-I	0.47	0.51	0.52	0.42	0.48	0.43	0.47	0.46	0.46	0.46
Q2.5-VL-7B-I	0.54	0.48	0.43	0.46	0.51	0.50	0.54	0.51	0.48	0.51
Q2.5-VL-32B-I	0.45	0.60	0.51	0.51	0.50	0.49	0.50	0.55	0.40	0.50
Q2.5-VL-72B-I	0.40	0.57	0.52	0.39	0.46	0.40	0.47	0.55	0.56	0.48
Q3-VL-4B-I	0.42	0.42	0.47	0.47	0.43	0.48	0.50	0.53	0.54	0.51
Q3-VL-8B-I	0.44	0.51	0.51	0.38	0.46	0.45	0.49	0.51	0.53	0.49
Q3-VL-32B-I	0.38	0.52	0.54	0.58	0.45	0.36	0.48	0.59	0.63	0.50
Q3-VL-8B-T	0.45	0.51	0.54	0.45	0.48	0.49	0.45	0.56	0.51	0.50
GLM-4.1V-9B-T	0.37	0.54	0.49	0.53	0.44	0.33	0.41	0.48	0.50	0.42
GPT-4o	0.40	0.63	0.50	0.56	0.48	0.38	0.43	0.50	0.53	0.45
GPT-5	0.66	0.63	0.77	0.58	0.66	0.62	0.66	0.64	0.66	0.64
Human Annotators	0.88	0.84	0.92	0.92	0.91	0.88	0.84	0.92	0.92	0.91
Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Table 2: Macro-F1 results on `VRRPI-Bench` (for short: Lm - Llama, Lv - Llava, Q - Qwen, I - Instruct, T - Thinking, SQ - SpaceQwen), comparing CV method (top), VLM (mid) and human (bottom). Note that, for human evaluation, we test on a mixed subsets from both datasets, hence we duplicate the results here. Red represents the best score, while blue shows the best score of VLMs. See App. C.1 for case study.

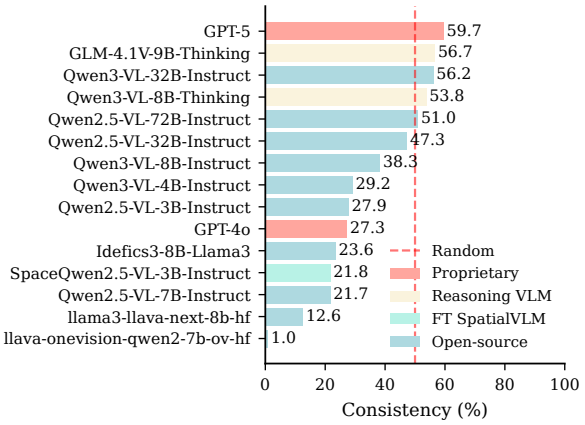

Figure 4: Consistency performance. The dashed line denotes the random baseline (50.0%), highlighting that most of VLMs are around or below random level.

5.2 Consistency Analysis

Previous research (Zheng et al., 2023; Qiu et al., 2024; Pezeshkpour and Hruschka, 2024) has established that LLMs are sensitive to the position of information in a prompt. We follow this observation in the multimodal domain and check if VLMs are stable when the temporal or logical order of vi-

sual inputs is permuted. We randomly sample 300 instances from VRRPI-Bench and swap the order of the two input images, treating the original source viewpoint as the target and vice versa, while keeping the textual prompt identical (the answer would be reversed). A model is considered consistent if it predicts the logically opposite option after the swap, regardless of the correctness of its initial prediction. As shown in Figure 4, most VLMs fail to maintain this basic logical symmetry. Even GPT-5, the top-performing model, achieves only 59.7% consistency. This significant failure rate suggests that the internal visual representations in current VLMs are not robustly aligned with the language space, leading to a breakdown in the spatial-order information preserved between images.[5]

Insight

VLMs fail to internalize the logical symmetry of image pairs, indicating that their visual representations do not robustly encode spatial-order information relative to the language prompt.

5.3 Diagnostic Analysis by VRRPI-Diag

The task in VRRPI-Bench involves composite spatial reasoning, requiring models to infer both rotational and translational camera motions. To disentangle these components, we introduce VRRPI-Diag, evaluating a single DoF (a rotation or a translation), along one of the camera axes.

Figure 5 shows models have gains when reasoning over isolated DoFs, yet performance is non-uniform across different axes. While VLMs are relatively proficient at identifying x and y translations, they struggle with transformations along the z-axis. We hypothesize that this discrepancy stems from a reliance on 2D motion heuristics; for small movements, transformations in the xy-plane often manifest as simple horizontal or vertical pixel shifts in the image plane, which models may track using low-level visual correspondences.

In contrast, z-axis transformations, specifically depth (translation) and roll (rotation), require a deeper internalization of projective geometry. Translation along the z-axis induces "looming" effects and scale changes, while roll alters the canonical orientation of the scene; neither can be approximated by simple 2D translations. This DoF-specific performance collapse suggests that VLMs rely on "pseudo-spatial" reasoning that fails when

[5]Input order was validated as non-issue, but performance collapsed on 3D scenes and our high-level task.

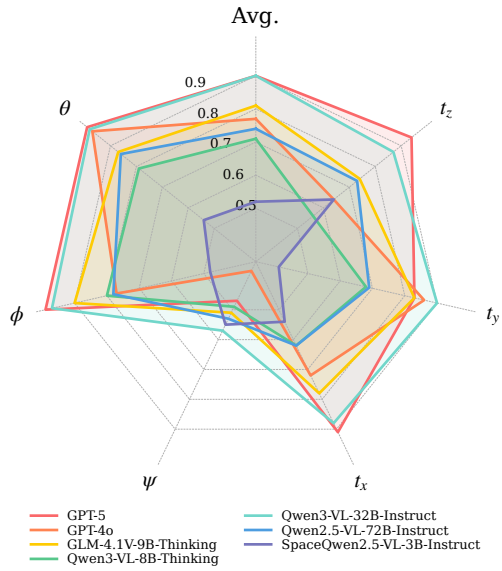

Figure 5: Results on VRRPI-Diag (top-performing VLMs of each category). The breakdown reveals a significant performance gap between image-plane shifts and optical-axis transformations. Full table in App. C.2.

confronted with the unique projective distortions associated with the optical axis.

Insight

Testing isolated DoFs reveals a dual failure: VLMs suffer from a "compositional gap" when multiple motions are combined, yet their gains on easier datasets are largely driven by 2D image-plane heuristics that fail to generalize to the z-axis (depth and roll).

6 Error Analysis

We use VRRPI-Diag to address three key questions: (i) Can VLMs recognize basic spatial relations within a single image (§6.1)? (ii) Can VLMs compare two images to ground the *same* object(s) and detect relational changes (§6.2)? (iii) Can VLMs improve in inferring the underlying camera-perspective motion when provided with textual descriptions (§6.3)?

6.1 Intra-image Spatial Position Understanding

To ensure that the failures with VRRPI-Bench are not merely due to lack of single-image spatial understanding, we revisit the findings of Kamath et al. (2023), who showed that state-of-the-art models as of 2023 struggled to reason about even simple spatial relations within a single image. To verify that this limitation does not persist in current VLMs,

	Whats-Up	COCO-spatial	GQA-spatial	Avg.	**One-obj**	**Two-obj**
BLIP2-ITM	37.6	53.0	49.8	46.8	-	-
BLIP2-ITC	29.0	53.7	51.0	44.6	-	-
Qwen2.5-VL-3B-Instruct	81.5	90.0	90.8	87.4	74.5	42.7
Qwen2.5-VL-32B-Instruct	92.9	92.5	91.8	92.5	73.5	70.7
Qwen3-VL-4B-Instruct	98.7	85.5	88.9	91.0	80.3	66.3
Qwen3-VL-32B-Instruct	99.4	87.2	93.1	93.2	85.3	76.0
GPT-4o	72.8	63.3	61.2	65.7	64.8	55.8
GPT-5	99.5	96.8	97.9	98.1	84.8	87.0
Random / Text-only	25.0	50.0	50.0	41.7	25.0	16.7
Human Estimate	100.0	97.3	99.0	98.8	-	-

Table 3: Performance on the WhatsUp (left) and our annotated dataset (**bold col.** on the right). We compare frontier VLMs against previous state-of-the-art models (e.g., BLIP-2) to establish a baseline for single-image spatial competence. See full table in App. D.1.

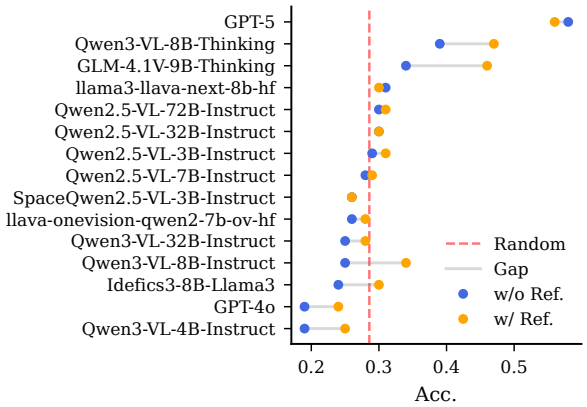

Figure 6: Cross-image relational comparison results. The dashed line denotes the random baseline (≈ 0.28). The bars represent the "Reference Gap", the performance gain achieved when the target object is explicitly identified in the prompt. See App. D.2 for full results.

explaining away their inability to perform more challenging tasks such as ours, we evaluate modern models on the Whatsup benchmark. This dataset focuses on 3D spatial relations (*left* and *right*) between objects within a single frame, where models are given image–caption pairs and must select the correct prepositional description from a candidate set. In addition, to ensure that our own data is not unsolvable due to similar low-level spatial ambiguities, we manually annotate a comparable single-image subset of `VRRPI-Diag`. Following their setup, we construct one-object tasks (determining an object's position within the image: *left, right, top, bottom*) and two-object tasks (determining the relative position between obj.1 and obj.2: *left, right, above, below, front, behind*) over images. Results on WhatsUp are shown on the left side of Table 3, while results on our annotated `VRRPI-Diag` subset appear on the right.

Newer models do significantly better. While previous state-of-the-art models like BLIP-2 struggled with these fundamental relations, current frontier models such as `GPT-5` and the larger variants of Qwen, demonstrate near-perfect performance, effectively reaching the ceiling of the WhatsUp benchmark. This suggests that the basic ability to resolve spatial prepositions within a static 2D context is largely a solved problem for modern VLMs. Consequently, the performance degradation observed in `VRRPI-Bench` cannot be attributed to a lack of basic spatial concepts, but rather to the inherent difficulty of 3D geometric reasoning across viewpoints.

Insight

Modern VLMs have achieved near-saturation on single-image spatial benchmarks, significantly outperforming earlier models like BLIP-2. This confirms that their primary limitation lies not in basic spatial vocabulary, but in the higher-order geometric reasoning required for multi-view consistency.

6.2 Cross-image Relational Comparison

Building on the intra-image baseline, we evaluate whether VLMs can maintain spatial consistency when comparing object relations across two disparate viewpoints. This task examines the model's ability to identify how the relative position of a shared reference object changes under camera motion. Given that 3D transformations can manifest as ambiguous visual shifts in 2D, we manually annotate each sample in `VRRPI-Diag` with all valid directional labels (e.g., *left, right, upwards, closer*). A prediction is deemed correct if it matches any ground-truth label in the set.

To isolate the source of error, we use two prompting strategies: (a) *w/o Ref.*—the model must autonomously identify a reference object and infer its displacement; (b) *w/ Ref.*—the reference object is explicitly provided, focusing the model's attention solely on the positional change.

Figure 6 reveals a systemic struggle with cross-view correspondence. While `GPT-5` defines the current frontier with near-saturated performance, the majority of models, including `GPT-4o`, fail to significantly exceed the random baseline. A "Think-

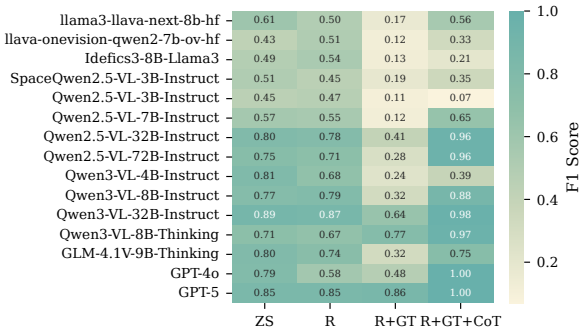

Figure 7: Camera-perspective reasoning. We evaluate whether models can synthesize object-level grounding into a global viewpoint transformation through four iterative setups: Zero-Shot (ZS), Ref Object Specification (R), Motion Grounding (R+GT), and Full Reasoning (R+GT+CoT).

ing" tier of models (e.g., Qwen3-VL-8B-Thinking) exhibits the largest performance gains when provided with an explicit reference. This indicates that for mid-tier models, the primary bottleneck is *cross-view object correspondence*, the ability to recognize the same object across different projections. However, even with this assistance, most models remain far below human-level consistency, suggesting that the underlying geometric mapping remains fundamentally unresolved.

Insight

While models like `GPT-5` begin to bridge the cross-view gap, most VLMs fail at the geometric mapping stage. The significant "Reference Gap" in reasoning-heavy models suggests that while they can resolve spatial changes, they lack the robust visual correspondence needed to autonomously link objects across viewpoints.

6.3 Camera-perspective Reasoning

Building on the previous diagnostic stages, we perform a sanity check to determine if the observed failures are rooted in a lack of perspective-transformation logic. We evaluate a subset of `VRRPI-Diag` under four iterative prompting setups: (1) **ZS**—the zero-shot baseline; (2) **R**—manually provided reference object identities; (3) **R+GT**—explicit object-motion annotations; and (4) **R+GT+CoT**—a complete reasoning chain describing the inverse relationship between object and camera motion.

In Figure 7, several patterns emerge. First, providing object-level grounding (*R*) fails to significantly improve performance, indicating that simply identifying relevant objects is insufficient for viewpoint reasoning. Second, the performance drop in *R+GT* shows that despite being told exactly how an object moved, most VLMs fail to recognize that if a shared object appears to move in one direction across views, the camera must have moved in the opposite direction.

Finally, the *R+GT+CoT* setup acts as a critical control. While frontier models like `GPT-5` reach near-perfect F1 scores when the full reasoning rule is explicitly stated in the prompt, this success appears to be a result of textual logic rather than visual understanding. By reducing the problem to a linguistic task, strong models can bypass the visual complexity. In contrast, smaller models are often distracted by the increased prompt complexity, leading to higher variance in their predictions.

We speculate that part of the challenge our benchmark poses to VLMs (versus the classic methods) is inherent in their architecture: they rely on coarse tokens to represent patches in the images, unaligned with the fine-grained nature of some of the tasks we present in our benchmark. This calls for future work in VLMs, where models become both sensitive to semantic, global information as well as low level (for example, at the pixel level) information.

Insight

The failure in the *R+GT* setup confirms a specific reasoning deficit: VLMs struggle to internalize the inverse relationship between object-level visual shifts and global camera motion. While explicit linguistic rules (*R+GT+CoT*) can compensate for this, it highlights a disconnect between the models' textual reasoning and their internal visual representations.

7 Conclusion

We introduced `VRRPI-Bench`, a benchmark for realistic multi-view 3D spatial reasoning. Most of VLMs are near random on the main benchmark. Our analyses show that this gap is not due to weak single-image spatial competence, but to failures in cross-view correspondence, view consistency, and projective camera-motion reasoning, especially for optical-axis transformations. We therefore view `VRRPI-Bench` not only as a challenging benchmark, but as a targeted diagnostic for the multi-view capabilities current VLMs still lack. Progress on this task would provide stronger evidence of grounded, view-consistent spatial reasoning in VLMs.

Limitations

First, our benchmark focuses on indoor RGB-D environments, constructed from existing 3D datasets. While this design enables precise pose computation and controlled evaluation, the exploration in diverse environments including outdoor scenes or dynamic environments is limited. Second, our study evaluates general-purpose VLMs in zero-shot setup. A potential direction is to investigate whether training objectives, such as geometric supervision or multi-view consistency constraints, can improve models' ability to reason about camera motion.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2025. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE.

Lorenzo Pasqualetto Cassinis, Robert Fonod, and Eberhard Gill. 2019. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft. *Progress in Aerospace Sciences*, 110:100548.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. 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.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2023. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36:70115–70140.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355.

Ali Elqursh and Ahmed Elgammal. 2011. Line-based relative pose estimation. In *CVPR 2011*, pages 3049–3056. IEEE.

Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Dinh-Cuong Hoang, Achim J Lilienthal, and Todor Stoyanov. 2020. Object-rpe: Dense 3d reconstruction and pose estimation with convolutional neural networks. *Robotics and Autonomous Systems*, 133:103632.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. Association for Computational Linguistics.

Jed M Kelsey, Jeffrey Byrne, Martin Cosgrove, Sanjeev Seereeram, and Raman K Mehra. 2006. Vision-based relative pose estimation for autonomous rendezvous and docking. In *2006 IEEE aerospace conference*, pages 20–pp. IEEE.

Fadi Khatib, Yuval Margalit, Meirav Galun, and Ronen Basri. 2024. Leveraging image matching toward end-to-end relative camera pose regression. *Preprint*, arXiv:2211.14950.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17028–17047, Miami, Florida, USA. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. 2025a. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*.

Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, and 1 others. 2025b. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. 2024. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *arXiv preprint arXiv:2403.13438*.

Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. 2015. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651.

Etienne Mouragnon, Maxime Lhuillier, Michel Dhome, Fabien Dekeyser, and Patrick Sayd. 2006. Real time localization and 3d reconstruction. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 363–370. IEEE.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Sarah Price. 2004. Expressing pose: rotation and translation matrices and vectors.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Prateeth Rao and 1 others. 2026. Relational epipolar graphs for robust relative camera pose estimation. *arXiv preprint arXiv:2604.04554*.

Chris Rockwell, Justin Johnson, and David F Fouhey. 2022. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE.

Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F. Fouhey. 2024. Far: Flexible, accurate and robust 6dof relative camera pose estimation. In *CVPR*.

Dhruv Shah, Błażej Osiński, Sergey Levine, and 1 others. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. 2013. Scene coordinate regression forests for camera relocalization in rgb-d images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2025. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780.

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

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2025. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4083–4103, Suzhou, China. Association for Computational Linguistics.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2025a. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025b. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, and 1 others. 2025. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

Xun S Zhou and Stergios I Roumeliotis. 2008. Robot-to-robot relative pose estimation from range measurements. *IEEE Transactions on Robotics*, 24(6):1379–1393.

Appendix

The list of appendices is as follows:

- Data collection (App. A): Details about algorithm we use for creating datasets.
- Experiment details (App. B): Details about prompts and how we evaluate computer vision methods, VLMs and human annotators.
- Main results (App. C): Case study on `VRRPI-Bench` and full table result for `VRRPI-Diag`.
- Error analysis (App. D): Full results on the WhatsUp dataset, together with visualization on our annotated dataset, and a full table for the cross-images grounding task.

A Data Collection

A.1 Geometry Computation

Below, we describe how to compute the relative pose vector $\boldsymbol{v}$ given two frames and their corresponding pose matrices, along with the angle τ and the central point deviation d between the two viewpoints and their central intersection point.

Camera Coordinate System We adopt the standard right-handed camera coordinate system, where the camera's optical center is located at the origin. The z-axis points forward, the x-axis points to the right, and the y-axis points downward.

Solve Relative Pose Vector from Extrinsic Parameters Suppose extrinsic matrix for i-, and j-th camera system, $\mathbf{T}_{wc_i}$ and $\mathbf{T}_{wc_j}$, are given. If we want to determine how the i-th camera moves to the j-th camera, we aim to express the origin of the j-th camera's coordinate system in terms of the i-th camera's coordinate system. This can be achieved through a geometric transformation:

$$\mathbf{T}_{c_ic_j} = (\mathbf{T}_{wc_i})^{-1}\mathbf{T}_{wc_j}, \tag{1}$$

where we can decompose $\mathbf{T}_{c_ic_j}$ into rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. The matrix $\mathbf{R}$ describes the rotation from the i-th camera to the j-th camera, and $\mathbf{t}$ represents the translation from the i-th camera to the j-th camera.

To obtain the relative pose vector $\boldsymbol{v}$, we use the following formulas from Price (2004):

$$\theta = \arctan\left(\frac{\mathbf{R}_{32}}{\mathbf{R}_{33}}\right), \tag{2}$$

$$\phi = \arcsin\left(-\mathbf{R}_{31}\right), \tag{3}$$

$$\psi = \arctan\left(\frac{\mathbf{R}_{21}}{\mathbf{R}_{11}}\right). \tag{4}$$

Thus, $\boldsymbol{v}$ fully characterizes how the i-th camera moves relative to the j-th camera in absolute terms.

Unprojection and Reprojection Unprojection refers to transforming a 2D coordinate into a 3D coordinate, while reprojection involves transforming a 3D coordinate into a 2D coordinate.

Angle (τ) and Deviation Between Two Perspectives After Reprojection Suppose we are given $\mathbf{T}_{wc}$ and $\mathbf{K}$ for the i-th and j-th cameras, along with the depth image D. Let $\boldsymbol{p}_i$ be the central point of the i-th image. To unproject it into the world coordinate, we use the following equation:

$$\boldsymbol{p}_w = D_{\boldsymbol{p}_i}\mathbf{T}_{wc_i}(\mathbf{K}_i)^{-1}\boldsymbol{p}_i, \tag{5}$$

where $D_{\boldsymbol{p}_i}$ denotes the depth value of the central point $\boldsymbol{p}_i$.

Next, we reproject this point onto the j-th image:

$$\boldsymbol{p}_j = \mathbf{K}_j(\mathbf{T}_{wc_j})^{-1}\boldsymbol{p}_w. \tag{6}$$

To compute the angle τ between the two viewpoints and the deviation d of the central point after reprojection, we first decompose the origins of the i-th and j-th cameras, $\mathbf{o}_i$ and $\mathbf{o}_j$, in the world coordinate system using $\mathbf{T}_{wc_i}$ and $\mathbf{T}_{wc_j}$. We then calculate:

$$\tau = \arccos\left(\frac{\vec{\boldsymbol{p}_w\mathbf{o}_i}\cdot\vec{\boldsymbol{p}_w\mathbf{o}_j}}{\|\vec{\boldsymbol{p}_w\mathbf{o}_i}\|\|\vec{\boldsymbol{p}_w\mathbf{o}_j}\|}\right), \tag{7}$$

$$d = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|. \tag{8}$$

Algorithm 1 Abstract generation of `VRRPI-Bench`.

Require: $\{(I_i, D_i, \mathbf{K}_i, \mathbf{T}_{wc_i})\}_{i=1,\dots,n}, \tau^-, d^+, l^-, l^+$.
Ensure: $I_i, I_j, \boldsymbol{v}^*, \tau^*, \bar{d}^*$

```
for i = 1, ..., n − 1 do
    for j = i + l^−, ..., min(n, i + l^+) do
        Calculate τ and d̄ in reprojection.
        if τ ∈ [τ^−, τ^− + 5] and d̄ < d^+ then
            T_{c_i c_j} ← (T_{wc_i})^{−1} T_{wc_j}.
            Decompose R, t from T_{c_i c_j}.
            Solve v from R, t.
            τ^* ← τ.
            d̄^* ← d̄.
        end if
    end for
end for
```

	δ^-	δ^+
θ	5	15
ϕ	5	15
ψ	3	10
t_x	0.15	0.4
t_y	0.1	0.3
t_z	0.15	0.4

Table 4: Thresholds used in `VRRPI-Diag`

A.2 Details of Algorithm and Pseudocode

`VRRPI-Bench` Algorithm 1 presents the details of our filtering process for `VRRPI-Bench`, where l^- and l^+ represent the minimal and maximal frame intervals, respectively, to make the filtering more efficient. In this process, we define an angular range τ^- and select four levels (15, 30, 45, 60) to simulate progressively more challenging cases. The predefined maximal central point deviation $\bar{d}$ is set to 300, as most images have a width of 640, and 300 is less than half of that.

`VRRPI-Diag` Algorithm 2 presents the details of our filtering process for `VRRPI-Diag`. To filter images that meet our requirements, we select meaningful thresholds that allow the images to show differences while still being part of the same scene. We use the following set of thresholds $\{(\delta_i^-, \delta_i^+)\}_{i=1,\dots,6}$, as shown in Table 4.

A.3 Dataset Statistics

Table 5 gives the statistics about `VRRPI-Diag`. Table 6 compares the general spatial reasoning benchmark with ours, highlighting our challenging setup.

Algorithm 2 Abstract generation of VRRPI-Diag.

Require: $\{(I_i, \mathbf{T}_{wc_i})\}_{i=1,\dots,n}, \{(\delta_i^-, \delta_i^+)\}_{i=1,\dots,6}, l^-, l^+$.
Ensure: $I_i, I_j, \boldsymbol{v}^*$.

```
for i = 1, ..., n − 1 do
    for j = i + l^−, ..., min(n, i + l^+) do
        T_{c_i c_j} ← (T_{w c_i})^{−1} T_{w c_j}.
        Decompose R, t from T_{c_i c_j}.
        Solve v from R, t.
        for k = 1, ..., 6 do
            if |v_k| > δ_k^+ then
                if |v_{m≠k}| < δ_m^− then
                    v* ← v.
                end if
            end if
        end for
    end for
end for
```

	DoF					
	θ	ϕ	ψ	t_x	t_y	t_z
# Frames	285	123	258	143	29	53

Table 5: Statistics for VRRPI-Diag.

	3D			Mental		Camera			Multi-view	
	RU	ME	CG	PP	SR	Rot.	Tr.	Comb.	Cont.	Disc.
WhatsUp	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
BLINK	✓	✗	✗	✗	✓	✗	✓	✗	✓	✓
Spatial-MM	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
SpatialEval	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Q-Spatial	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
SpatialRGPT	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗
MM-Spatial	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗
RoboSpatial	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
VSI-Bench	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗
MindCube	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
LRR-Bench	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
VRRPI-Bench	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
VRRPI-Diag	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗

Table 6: Comparison of spatial reasoning benchmarks by task categories (RU - Relation Understanding, ME - Metric Estimation, CG - Coordinate Grounding, PP - Path Plan, SR - Scene Reasoning, Rot. - Rotation, Tr. - Translation, Comb. - Combination, Cont. - Continuous, Disc. - Discrete), highlighting the novelty and focus of our work on spatial reasoning research.

B Experimental Setup

B.1 Baseline Evaluation

We use SIFT and LoFTR to detect matching points across images, then apply RANSAC to compute the relative pose, which is represented in homogeneous pose matrix format. From this, we extract the rotation matrix and translation vector. The relative pose vector $\boldsymbol{v}$ is then obtained, and we check the sign of the translation along the y-axis. This is because we aim to maintain the same label space as in VLM—binary classification and translation direction is the discriminative variable that uniquely determines the paired binary label under our dataset construction. Therefore, we evaluate the performance of the computer vision methods based on the sign of the y-axis translation in their predicted values.

B.2 VLM Evaluation

To enhance the reasoning ability of VLMs, we use a special output format instead of short-output. Specifically, we want the VLM to reason first within the <thinking> tags and then provide the answer within the <ans> tags. To minimize bias, we also randomly shuffle the order of answer candidates for each sample.

Open-souce VLMs are evaluated using the HuggingFace Transformers library (Wolf et al., 2020), while closed-source VLMs are evaluated through their respective APIs. We use greedy decoding for all models to ensure that they strictly follow the output format. Our experiments are conducted on a single (or multi-GPUs if needed) NVIDIA H200 GPU with 100GB memory.

The example prompts we use for VRRPI-Bench and for VRRPI-Diag can be found in Figure 8 and Figure 9, respectively.

B.3 Human Evaluation

We randomly select samples from VRRPI-Bench and design a Google Form to collect data. The same prompt is used for human annotators, with the exception of special tokens and formatting, to minimize any potential bias.

The prompt is then presented to human annotators (see Figure 10), with the options provided in the Google Form.

C Main Results

C.1 Case Study on VRRPI-Bench

As shown in Figure 11, GPT-5 accurately detects local objects, yet fails to describe their relative shifts and infer the correct camera motion between viewpoints rigorously. This confusion stems from the inherent ambiguity between camera rotation and translation, revealing a deeper misalignment between visual perception and 3D spatial reasoning.

C.2 Full Results on VRRPI-Diag

Table 7 shows the full numbers for Figure 5.

Example prompt for VLM on VRRPI-Bench

```
<input>
You are given two consecutive frames from a scene. The first image shows the
    **source viewpoint**, and the second image shows the **target viewpoint**. The
    camera moves and rotates simultaneously between these two frames. Use the
    visual difference between them to reason about the camera motion in 3D space.
</input>

<task>
Select the correct description of the camera motion between these viewpoints.
</task>

<ans-candidates>
0. Move left while yawing right
1. Move right while yawing left
</ans-candidates>

<output-format>
Please strictly follow the format: Provide your spatial reasoning inside
    <thinking></thinking> XML tags, and provide **only index of your option**
    inside <ans></ans> XML tags, e.g., <thinking>...</thinking> <ans>...</ans>.

Do not output anything else.
</output-format>
```

Figure 8: Example prompt for VLM on VRRPI-Bench.

D Error Analysis

D.1 Full Results and Visualization for Intra-Image Analysis

We show the full table of VLMs on WhatsUp dataset in Table 8. Also, Figure 12 present the visualization about how VLMs perform on our annotated dataset, splitted by each spatial relation.

D.2 Full Results for Cross-Image Analysis

Table 9 shows the full numbers for Figure 6.

Example prompt for VLM on VRRPI-Diag

```
<input>
You are given two consecutive frames from a scene. The first image shows the
    **source viewpoint**, and the second image shows the **target viewpoint**. The
    camera movement is usually described in 6 degree of freedom
    (6DoF)---translation and rotation along x, y, and z axes. We now have a
    constraint on camera motion between source and target viewpoints, that is, the
    significant movement is only rotation along y-axis (yaw), i.e., rotate left or
    right.
</input>

<task>
Use the visual difference between them to reason about the direction of this
    significant camera motion in 3D space. Select the correct description of the
    camera motion between these viewpoints.
</task>

<ans-candidates>
0. Rotate right
1. Rotate left
</ans-candidates>

<output-format>
Please strictly follow the format: Provide your spatial reasoning inside
    <thinking></thinking> XML tags, and provide **only index of your option**
    inside <ans></ans> XML tags, e.g., <thinking>...</thinking> <ans>...</ans>.

Do not output anything else.
</output-format>
```

Figure 9: Example prompt for VLM on VRRPI-Diag.

Prompt for human annotators

```
You are given two consecutive frames from a scene. The first image shows the
    source viewpoint, and the second image shows the target viewpoint. The camera
    moves and rotates simultaneously between these two frames. Use the visual
    difference between them to reason about the camera motion in 3D space.
```

Figure 10: Prompt for human annotators.

	θ	ϕ	ψ	t_x	t_y	t_z	Avg.
llama3-llava-next-8b-hf	0.59	0.55	0.52	0.61	0.38	0.77	0.58
llava-onevision-qwen2-7b-ov-hf	0.46	0.32	0.54	0.38	0.37	0.37	0.42
Idefics3-8B-Llama3	0.42	0.58	0.48	0.54	0.67	0.57	0.52
SpaceQwen2.5-VL-3B-Instruct	0.54	0.48	0.55	0.54	0.41	0.64	0.52
Qwen2.5-VL-3B-Instruct	0.51	0.54	0.52	0.61	0.46	0.70	0.55
Qwen2.5-VL-7B-Instruct	0.59	0.60	0.49	0.60	0.73	0.67	0.59
Qwen2.5-VL-32B-Instruct	0.86	0.78	0.51	0.64	0.76	0.69	0.74
Qwen2.5-VL-72B-Instruct	0.86	0.78	0.53	0.62	0.69	0.73	0.74
Qwen3-VL-4B-Instruct	0.78	0.73	0.48	0.68	0.69	0.67	0.70
Qwen3-VL-8B-Instruct	0.68	0.84	0.49	0.75	0.63	0.74	0.72
Qwen3-VL-32B-Instruct	0.98	0.97	0.57	0.88	0.90	0.87	0.90
Qwen3-VL-8B-Thinking	0.79	0.80	0.49	0.62	0.68	0.56	0.71
GLM-4.1V-9B-Thinking	0.87	0.90	0.51	0.78	0.83	0.74	0.81
GPT-4o	0.97	0.77	0.37	0.72	0.86	0.64	0.77
GPT-5	0.99	0.99	0.47	0.91	0.83	0.94	0.90

Table 7: Full table of results on VRRPI-Diag. See Figure 5 for visualization about top-performing VLMs.

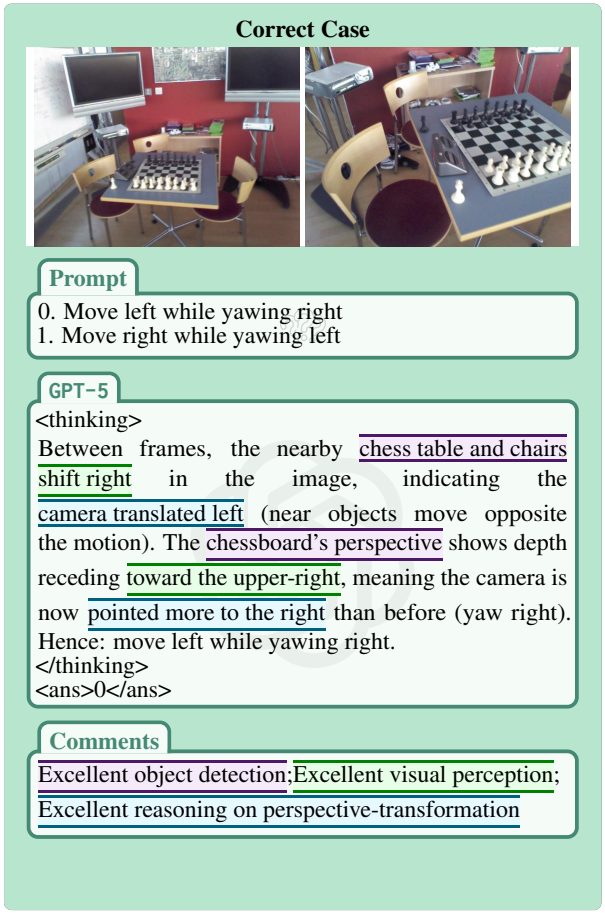

Figure 11: Qualitative examples from `GPT-5` on `VRRPI-Bench`. Although the model correctly identifies objects, it fails to ground relative shifts and reason about the underlying 3D camera motion (right), revealing the inherent ambiguity between rotation and translation.

	What'sUp Subset A			**What'sUp Subset B**			**COCO-spatial**		**GQA-spatial**		
	Indiv.	**Pairs**	**Set of 4**	**Indiv.**	**Pairs**	**Set of 4**	**One-obj**	**Two-obj**	**One-obj**	**Two-obj**	**Avg.**
BLIP2-ITM	44.9	24.3	0.0	30.4	2.0	0.0	48.3	57.7	46.0	53.6	46.8
BLIP2-ITC	35.9	3.4	0.0	22.1	0.0	0.0	55.6	51.8	52.6	49.5	44.6
Idefics3-8B-Llama3	53.9	36.4	2.9	51.0	27.9	5.9	75.9	67.0	78.8	69.4	66.0
llama3-llava-next-8b-hf	25.5	4.3	0.0	22.3	3.9	0.0	48.6	53.0	48.8	57.4	42.6
llava-onevision-qwen2-7b-ov-hf	91.0	84.5	70.9	74.0	56.9	25.5	85.0	82.5	87.5	87.3	84.5
SpaceQwen2.5-VL-3B-Instruct	83.3	69.4	44.7	71.6	23.1	1.9	68.2	65.7	68.2	67.4	70.7
Qwen2.5-VL-3B-Instruct	89.3	82.0	66.0	73.8	60.3	27.5	92.2	87.7	93.0	88.7	87.4
Qwen2.5-VL-7B-Instruct	86.2	76.7	53.4	83.6	71.1	49.0	94.0	89.8	94.3	83.2	88.5
Qwen2.5-VL-32B-Instruct	98.1	96.6	93.2	87.7	79.9	59.8	93.3	91.8	94.7	89.0	92.5
Qwen2.5-VL-72B-Instruct	99.0	98.5	97.1	95.6	91.7	83.3	93.9	96.1	96.0	95.2	96.2
Qwen3-VL-4B-Instruct	99.5	99.5	99.0	97.8	95.6	91.2	89.0	82.0	90.6	87.3	91.0
Qwen3-VL-8B-Instruct	96.6	93.7	87.4	95.3	91.2	82.4	93.6	83.9	95.1	89.0	92.3
Qwen3-VL-32B-Instruct	99.5	99.5	99.0	99.3	98.5	97.1	89.6	84.8	94.4	91.8	93.2
GLM-4.1V-9B-Thinking	99.0	98.5	97.1	96.8	94.1	88.2	96.2	95.9	97.2	98.6	97.3
GPT-4o	82.8	70.9	48.5	62.7	35.3	5.9	62.8	63.9	65.9	56.4	65.7
GPT-5	99.5	99.5	99.0	99.5	99.0	98.0	96.8	96.8	97.3	98.6	98.1
Random chance	25.0	6.3	0.4	25.0	6.3	0.4	50.0	50.0	50.0	50.0	41.7

Table 8: Performance on the `WhatsUp` dataset (following their metrics). We compare current frontier VLMs against previous state-of-the-art models (e.g., BLIP-2) to establish a baseline for single-image spatial competence.

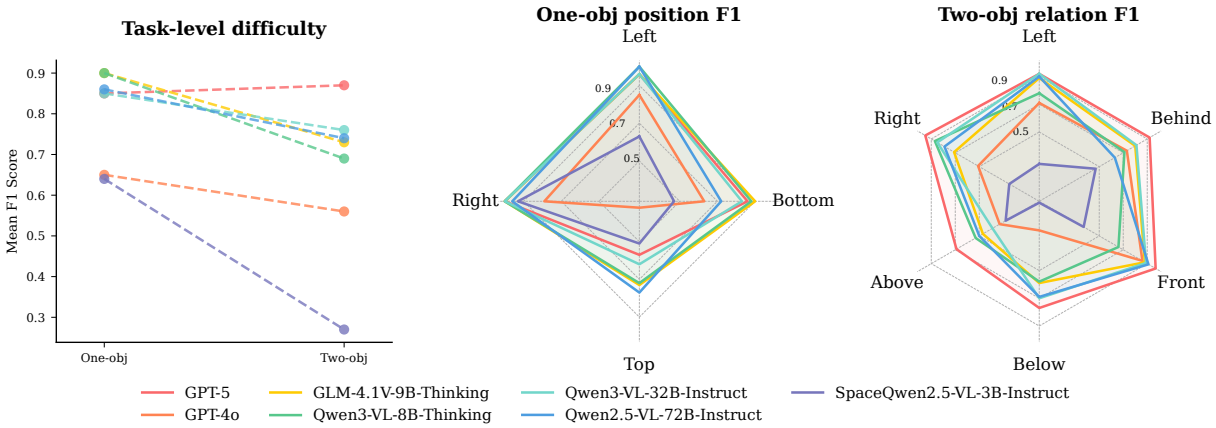

Figure 12: Evaluating intra-image spatial reasoning. This stage ensures that models possess the requisite spatial vocabulary before moving to multi-view geometric tasks.

	w/o Ref.	w/ Ref.
llama3-llava-next-8b-hf	0.31	0.30
llava-onevision-qwen2-7b-ov-hf	0.26	0.28
Idefics3-8B-Llama3	0.24	0.30
SpaceQwen2.5-VL-3B-Instruct	0.26	0.26
Qwen2.5-VL-3B-Instruct	0.29	0.31
Qwen2.5-VL-7B-Instruct	0.28	0.29
Qwen2.5-VL-32B-Instruct	0.30	0.30
Qwen2.5-VL-72B-Instruct	0.30	0.31
Qwen3-VL-4B-Instruct	0.19	0.25
Qwen3-VL-8B-Instruct	0.25	0.34
Qwen3-VL-32B-Instruct	0.25	0.28
Qwen3-VL-8B-Thinking	0.39	0.47
GLM-4.1V-9B-Thinking	0.34	0.46
GPT-4o	0.19	0.24
GPT-5	0.58	0.56

Table 9: Full table for cross-image task. See Figure 6 for visualization.