Improving Allocentric Spatial Reasoning of VLMs via Perspective-conditional Prompt-based Contrastive Decoding

Hyejin Chung **Hyunsoo Yoon**[†]
Yonsei University
{hjhj999, hs.yoon}@yonsei.ac.kr

Abstract

Vision-Language models (VLMs) have demonstrated impressive capabilities across diverse domains, yet they fail significantly in allocentric spatial reasoning, which requires perspective-taking from object-centric viewpoints rather than the camera's perspective. This capability is essential for real-world applications including robotics, autonomous driving, and embodied AI. In this work, we identify this limitation as a prevalent camera viewpoint bias and propose Perspective-Conditional Prompt-based Contrastive Decoding (PCD), a training-free method that leverages VLMs' capabilities while exploiting this performance gap. By contrasting responses to allocentric versus egocentric framings of identical spatial configurations, PCD explicitly suppresses camera viewpoint bias without requiring fine-tuning or external foundation models. Evaluating on allocentric benchmarks including 3DSRBench, OmniSpatial, ViewSpatial-Bench, and SpatialMQA, we demonstrate that PCD improves accuracy by up to 32.3 points on Qwen2.5-VL-7B, with consistent gains observed on recent, larger-scale VLMs. Our approach offers substantial computational advantages, achieving 8.65× faster inference and a 37.4% memory reduction compared to multi-model pipeline framework.

1 Introduction

With the advancement of Vision-Language Models (VLMs), they are being actively utilized across various domains, including autonomous driving, robotics, and beyond (Tian et al., 2025; Li et al., 2026; Song et al., 2025). A prominent area of research focuses on enhancing VLMs' spatial reasoning capabilities (Deitke et al., 2025; Chen et al., 2024; Cheng et al., 2024; Ma et al., 2026; Tong et al., 2024a).

Initial VLMs showed strong capabilities on Visual Question Answering (VQA) and image captioning (Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023), leading to widespread adoption and increasing demand for handling complex spatial reasoning tasks (Li et al., 2025a; Tong et al., 2024a; Ghosh et al., 2024; Tong et al., 2024b; Li et al., 2024; Liu et al., 2025b; Ma et al., 2025).

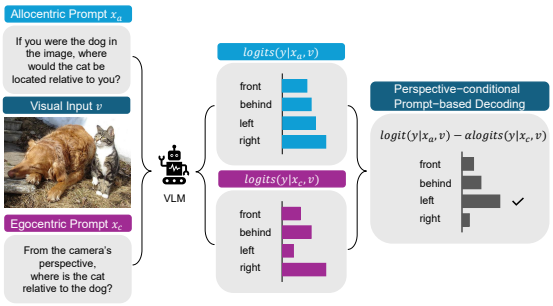


Figure 1: An illustration of Perspective-conditional Prompt-based Contrastive Decoding. The direction "right" is eliminated by contrasting with the output distribution that favors egocentric view. Therefore, the model can answer allocentric question properly.

To address these challenges, researchers have explored diverse approaches, ranging from instruction-tuning to architectures that integrate additional adapters for incorporating auxiliary inputs such as depth information (Chen et al., 2024; Cheng et al., 2024). However, VLMs still exhibit limitations (Jung et al., 2025; Zhang et al., 2025b; Kamath et al., 2023; Shiri et al., 2024). They show suboptimal performance even on spatial relation judgment tasks that are trivial for humans. Performance drops in allocentric spatial reasoning tasks, which require determining relative spatial relationships between objects from a perspective other than the camera's viewpoint (Góral et al.; Zhang et al., 2025a; Ma et al., 2025; Li et al., 2025b). This allocentric spatial reasoning is a critical capability for applications in robotics and autonomous driving (Chen et al., 2024; Cheng et al., 2024; Tian et al., 2025; Li et al., 2026).

[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21917–21935
October 24-29, 2026 ©2026 Association for Computational Linguistics

As noted in prior work (Zhang et al., 2025b; Liu et al., 2025a), this limitation stems from inconsistencies in training data, including annotation variability or cultural differences in spatial language conventions. Most existing research has focused on improving spatial reasoning from egocentric (camera) perspectives (Jung et al., 2025; Kamath et al., 2023), despite VLMs exhibiting substantially lower performance on allocentric tasks (Shiri et al., 2024; Ma et al., 2025; Lee et al., 2025). Consequently, while egocentric spatial reasoning has seen notable improvements (Cheng et al., 2024; Chen et al., 2024; Ma et al., 2026), allocentric perspective-taking remains a significant challenge. Even when models are explicitly instructed to adopt an allocentric viewpoint, they tend to respond based on the familiar camera perspective. We identify this persistent behavior as *camera viewpoint bias*, a limitation that affects allocentric spatial reasoning. This camera viewpoint bias is not merely a failure mode but a systematic pattern. VLMs consistently answer spatial questions from the camera's perspective even when explicitly instructed otherwise, achieving significantly higher accuracy on egocentric queries (Section 3.1).

In this work, we concentrate on allocentric spatial relationship (Lee et al., 2025), and propose PCD (Perspective-conditional Prompt-based Contrastive Decoding), a methodology that mitigates bias through contrastive decoding (Li et al., 2023; Leng et al., 2024), thereby improving the accuracy of allocentric spatial reasoning. Our approach builds on the observation that VLMs answer egocentric questions with higher accuracy compared to allocentric ones. By treating camera perspective judgments as a bias signal, we capture the model's response shift when allocentric perspective reasoning is required. Our approach leverages the VLM's own capabilities to extract objects relevant to the query and determine whether the viewer's perspective aligns with the camera's perspective.

Following the standard setup of most spatial relationship QA tasks (Ma et al., 2025; Jia et al., 2026; Li et al., 2025b; Wang et al., 2025b; Liu et al., 2024d), our method takes an image and a perspective-based question as input and processes them through 3 stages. First, we extract the reference object and the target object from the textual question, utilizing the VLM to obtain responses efficiently. Second, we identify whether the reference object's perspective aligns with the camera's viewpoint for spatial relationship judgment purposes, again leveraging the VLM to make this determination based on the appearance of the reference object. Finally, if the reference object's viewpoint aligns with the camera's perspective, the model generates answers directly. Otherwise, we generate answers by exploiting the contrast with camera perspective bias through contrastive decoding.

This methodology is training-free and does not require external models, offering advantages in terms of shorter inference time and lower memory overhead compared to other approach (Lee et al., 2025). We demonstrate superior performance over existing methods on multiple real-image-based spatial relation benchmarks, including 3DSRBench (Ma et al., 2025), OmniSpatial Benchmark (Jia et al., 2026), SpatialMQA (Liu et al., 2025a), and ViewSpatial-Bench (Li et al., 2025b).

Our finding is that camera viewpoint bias in VLMs is predominantly interpretive rather than perceptual (Sec. 3.4). Models correctly perceive spatial configurations but default to the camera viewpoint when interpreting the question. This diagnosis is what makes PCD an effective correction here, and is what distinguishes PCD's textual-framing contrast from prior CD variants that contrast model capacity or image perturbation. PCD is one practical, training-free instantiation of this diagnosis, targeted specifically at relative directional judgments in single-image allocentric perspective-taking, the setting adopted by existing allocentric benchmarks.

Our main contributions are:

- **Identification of egocentric bias as a major contributing factor:** We identify and formalize camera viewpoint bias as the primary factor limiting VLMs' allocentric spatial reasoning. We demonstrate that VLMs achieve high accuracy when questions align with camera perspective but fail on identical spatial relations from other viewpoints, isolating bias as the bottleneck.

- **Perspective-conditional Prompt-based Contrastive Decoding (PCD):** We propose a training-free method that factors out camera viewpoint bias by contrasting allocentric vs. egocentric question framings within a single VLM. Compared to prior work requiring external vision modules or fine-tuning, PCD achieves 37.41% memory reduction and 8.65× faster inference.

- **Substantial improvements across bench-**

marks: PCD achieves +32.3 points on 3DSRBench (72.29%), +31.0 on OmniSpatial (64.04%), and +23.4 on SpatialMQA (47.02%) with Qwen2.5-VL-7B, and +12.7 on ViewSpatial-Bench (54.63%) with InternVL2.5-8B, demonstrating generalization across VLM architectures (Qwen2.5-VL, Qwen3-VL, InternVL2.5, InternVL3.5, LLaVA-OneVision) and model scales (3B–38B parameters).

2 Related Work

2.1 Spatial Reasoning with VLMs

Despite recent advances, VLMs exhibit limitations in spatial reasoning, with state-of-the-art models performing at or below random chance on basic spatial tasks (Liu et al., 2023, 2024a; Li et al., 2022; Zhang et al., 2025b; Kamath et al., 2023; Liu et al., 2024d; Shiri et al., 2024; Ma et al., 2025; Liu et al., 2025a; Li et al., 2025b). Fine-tuning approaches leverage data collection and synthesis (Chen et al., 2024; Deitke et al., 2025), architectural innovation (Tong et al., 2024a), or task-specific training (Cheng et al., 2024), but provide limited allocentric coverage.

Training-free frameworks such as APC-Vis (Lee et al., 2025) convert allocentric queries using external vision modules, increasing inference overhead, while others (Liu et al., 2025a; Li et al., 2025b) construct large-scale datasets requiring substantial computational resources.

Prior works have identified pervasive egocentric bias in VLMs (Shiri et al., 2024; Zhang et al., 2025b; Góral et al.; Zhang et al., 2025a), where models exhibit strong preference for egocentric interpretations. Zhang et al. (2025b) evaluate frame-of-reference conversion using synthetic scenes, while Ma et al. (2025), Jia et al. (2026), and Li et al. (2025b) provide benchmarks specifically targeting allocentric spatial reasoning. However, existing approaches share limitations. Fine-tuning requires large-scale datasets and computational resources, while training-free methods rely on multiple foundation models, increasing memory overhead and inference latency. In contrast, our method directly addresses camera bias within a single VLM through PCD, achieving competitive performance without external dependencies or fine-tuning.

2.2 Contrastive Decoding for Vision-Language Models

Li et al. (2023) introduce contrastive decoding(CD), exploiting that failure modes are more prevalent in smaller models. CD contrasts an expert LM against an amateur by searching for text that maximizes their log-probability difference: $\log p_{\text{EXP}}(y_t|x, y_{<t}) - \log p_{\text{AMA}}(y_t|x, y_{<t})$. This training-free approach consistently outperforms nucleus sampling and typical decoding.

Leng et al. (2024) propose Visual Contrastive Decoding (VCD) to address object hallucinations in VLMs by contrasting distributions from original versus Gaussian-perturbed images: $\text{softmax}[(1+\alpha)\log p(y_t|v, x) - \alpha \log p(y_t|v', x)]$, where v' is the distorted image. However, VCD is designed for hallucinations rather than interpretation bias (perspective). VLMs accurately perceive spatial configurations from the camera's perspective but tend to default to egocentric interpretations even when allocentric reasoning is explicitly requested. Visual perturbation cannot address such biases and may disrupt fine-grained spatial cues (object orientation, pose) necessary for accurate spatial reasoning. Our empirical analysis (Appendix C.1.2) confirms VCD provides limited and inconsistent benefits for allocentric reasoning.

This gap follows directly from our diagnosis that camera viewpoint bias is primarily interpretive rather than perceptual (Sec. 3.4). Since the VLM already perceives the scene correctly and misapplies the camera's frame when interpreting the question, the appropriate contrastive signal is a textual reframing of the same question, not a change in model capacity (Li et al., 2023) or in the image itself (Leng et al., 2024). PCD operationalizes this diagnosis as a lightweight, training-free correction for this specific bias, rather than a general-purpose contrastive decoding method.

Our work bridges spatial reasoning and contrastive decoding by introducing Perspective-conditional Prompt-based Contrastive Decoding (PCD). Unlike prior training-free methods that require multiple foundation models (Lee et al., 2025) or operate on image space (Leng et al., 2024), PCD factors out egocentric bias through textual modifications within a single VLM. This offers three key advantages: (1) reduced memory and inference latency, (2) direct targeting of perspective bias, and (3) compatibility with any instruction-tuned VLM.

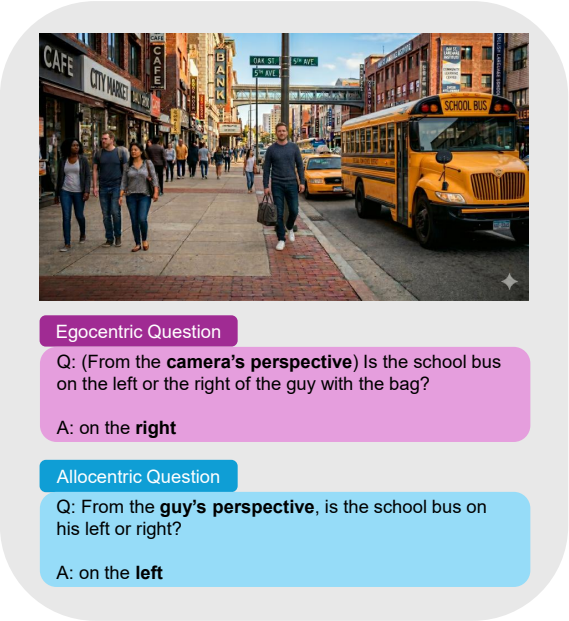


Figure 2: Comparison of egocentric and allocentric spatial reasoning questions. Even when presented with the same visual scene, spatial relationship queries yield different answers depending on the reference frame. The image is generated by Nano Banana 2 (Google, 2026).

3 Method

3.1 Overview

VLMs exhibit a performance asymmetry. They answer spatial questions more accurately when the reference frame aligns with the camera viewpoint. This asymmetry is not noise but signal. The higher egocentric accuracy captures the model's camera-centric bias. By subtracting this bias signal from allocentric reasoning, we can recover the underlying spatial understanding. We address allocentric spatial relationship reasoning in VLMs through a three-stage pipeline: (1) Object Extraction, (2) Viewpoint Alignment Verification, and (3) Perspective-Conditional Prompt-based Contrastive Decoding (PCD, Section 3.4). Since VLMs inherently approach spatial reasoning from an egocentric perspective (Zhang et al., 2025b), the model must identify the reference viewer to accurately determine relative spatial relations (Lee et al., 2025). We verify whether the camera view aligns with the reference viewer's egocentric viewpoint to determine the use of PCD. If aligned, the model can answer directly. Otherwise, camera perspective bias hinders the model's decision, so we suppress it through PCD to adopt the reference viewer's perspective.

Stage 1: Object Extraction. Given an allocentric question x_a and an image v, we extract reference object o_{ref} and target object o_{tgt}. o_{ref} is the object whose viewpoint defines the spatial frame of reference for determining relative positions. o_{tgt} is the object whose relative position is being queried. o_{ref} is then used in the subsequent stage, Viewpoint Alignment Verification, to assess whether the camera viewpoint aligns with o_{ref}'s perspective. We perform extraction via prompting to the same VLM(*e.g., Given question: "If I stand at the dog's position, is the ball on the left or right?" Reference object: dog, Target object: ball*). We use few-shot prompting for in-context learning and prompt templates are in Appendix B.2.

Stage 2: Viewpoint Alignment Verification. We determine whether the camera's viewpoint aligns with o_{ref}'s facing direction by prompting the model to identify which side of o_{ref} is visible. This approach proves more effective for vanilla VLMs than directly asking them for facing direction of o_{ref}. Based on the model's response, if the verification indicates that o_{ref}'s viewpoint is aligned with the camera viewpoint, we proceed with standard autoregressive generations. Otherwise, we apply PCD to handle the model's egocentric bias.

Viewpoint alignment verification is crucial for this task because contrastive decoding is not always beneficial (Suo et al., 2025). When the camera's viewpoint naturally aligns with the reference object's facing direction, the camera-centric answer is already correct for the allocentric question. In such cases, applying CD can degrade performance by suppressing the correct answer. Therefore, our viewpoint-conditional pipeline selectively applies PCD only when necessary. Moreover, it is the part of our pipeline as it enables both computational efficiency and improved accuracy.

Stage 3: Perspective-conditional Prompt-based Contrastive Decoding. Based on the result of viewpoint alignment verification, if the verification indicates that o_{ref}'s viewpoint is aligned with the camera viewpoint, we proceed with standard generation. Otherwise, we apply PCD to suppress camera viewpoint bias. This conditional approach avoids unnecessary CD application and reduces the risk of degrading performance on already-correct camera-centric answers.

Model	3DSRBench			OmniSpatial	SpatialMQA	ViewSpatial
	Overall	front/behind	left/right			
Random	50.00	50.00	50.00	25.00	25.59	25.00
Open-source VLMs						
LLaVA-1.5-7B	17.03	28.78	5.44	25.62	25.83	39.31
SpaceLLaVA	34.05	34.59	33.52	34.48	26.05	28.50
LLaVA-NeXT	37.81	41.57	34.10	40.39	26.05	18.41
LLaVA-OneVision	39.25	46.51	32.09	24.63	21.85	35.75
Molmo	21.36	22.97	19.77	39.90	28.26	29.93
Cambrian-1	45.31	49.71	40.97	26.11	33.77	35.04
Qwen2.5-VL-7B	39.97	45.06	34.96	33.00	23.62	34.92
InternVL2.5-8B	45.17	60.17	30.37	40.89	28.48	41.92
Proprietary Models						
GPT-5.4	32.03	43.90	20.34	40.89	34.20	35.27
Gemini-2.5-Pro	32.32	57.27	7.74	37.93	36.64	42.52
Grounded VLMs						
SpatialRGPT	26.58	27.19	25.98	34.31	25.39	34.68
Training-free VLM-in-the-loop Framework						
APC-Vis	70.13	67.44	72.78	18.72	33.41	39.19
Fine-tuned Model						
SpaceLLaVA LoRA	42.14	50.29	34.10	44.83	**51.66**	42.76
Ours						
PCD-Qwen2.5-VL-7B	**72.29**	**68.31**	**76.22**	**64.04**	47.02	45.61
PCD-InternVL2.5-8B	67.82	64.53	71.06	63.55	47.46	**54.63**

Table 1: Main results on allocentric spatial reasoning benchmarks (3DSRBench, OmniSpatial, SpatialMQA, ViewSpatial-Bench). Best results are in **bold**.

3.2 Decoding of Vision-Language Models

We consider a VLM parametrized by θ. Given an image input v where v provides essential visual context for generating an accurate response and an allocentric textual question x, a VLM generates the text answer y. Mathematically, it is formulated as:

$$y_t \sim p_\theta(y_t \mid v, x, y_{<t}) \propto \exp logit_\theta(y_t \mid v, x, y_{<t}) \quad (1)$$

where y_t denotes the token generated at time step t, and $y_{<t}$ represents the sequence of generated tokens up to time step $t-1$.

Prior works (Shiri et al., 2024; Zhang et al., 2025b; Góral et al.; Ma et al., 2025) have shown that VLMs exhibit strong camera-centric bias. They tend to answer spatial questions from the viewpoint of camera. This occurs because VLMs learn image-caption pairs describing scenes from the camera's perspective, creating a statistical bias in the model's learned distribution. Our goal is to suppress this camera-centric bias during inference without additional training or model modification.

3.3 Egocentric Camera Perspective Bias

Despite the prevalent camera perspective bias of VLMs, we observed cases where the model demonstrates flexibility and accuracy in spatial reasoning. In these cases, the model correctly answers spatial relationship questions by adopting the given reference object's perspective and also the camera viewpoint question (Figure 2). From this observation, we propose PCD, which explicitly leverages the perspective differences in textual prompts.

3.4 Perspective-conditional Prompt-Based Contrastive Decoding

We observe that VLMs respond properly to the same spatial reasoning question when framed from different viewpoints. While the model may incorrectly answer an allocentric question due to camera viewpoint bias, it can often accurately answer the spatial relationship question when explicitly asked from the camera's perspective. This observation suggests that the camera-centric bias is not purely a visual processing limitation, but rather a bias in how the model interprets and processes the textual question given its training distribution.

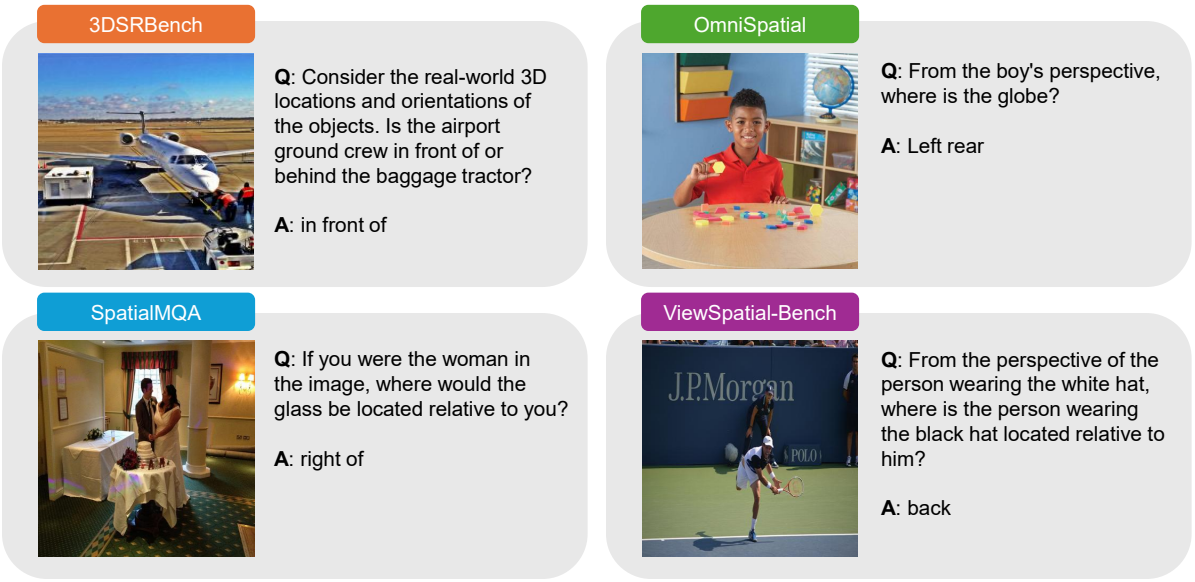


Figure 3: An illustration of spatial reasoning benchmarks (3DSRBench, OmniSpatial Benchmark, SpatialMQA, and ViewSpatial-Bench). The tasks probe spatial reasoning across relative directions.

Building on this insight, we introduce PCD. PCD counteracts egocentric camera perspective bias by contrasting model outputs generated from original allocentric text inputs and camera perspective textual prompts. This is achieved without requiring additional training or external pretrained models, making PCD a cost-effective and efficient approach.

Unlike VCD (Leng et al., 2024), which contrasts a clean image against a Gaussian-perturbed version, PCD keeps v unchanged and instead contrasts allocentric versus egocentric framings of the same given textual question. Our method is visualized in Figure 1.

Given the same image v, we construct two prompt variants: allocentric prompt and camera perspective prompt. Allocentric prompt x_{a} explicitly requests allocentric reasoning from the reference object's viewpoint(*e.g., "If I stand at the man's position facing where he is facing, is the car on the left or right of me?"*). Camera perspective prompt x_{c} explicitly requests camera-centric reasoning (*e.g., "From the camera's perspective, is the car on the left or right of the man?"*).

Our PCD is formulated as:

$$p_{\mathrm{PCD}}(y \mid v, x_{\mathrm{a}}, x_{\mathrm{c}}) = \\ \mathrm{softmax}\left[\mathrm{logit}_\theta(y \mid v, x_{\mathrm{a}}) \; -\alpha \cdot \mathrm{logit}_\theta(y \mid v, x_{\mathrm{c}})\right] \tag{2}$$

where larger α indicates a stronger bias suppression ($\alpha = 0$ reduces to regular decoding).

PCD serves as a corrective mechanism that amplifies the model's allocentric reasoning capability while suppressing its camera viewpoint bias. It can be interpreted as a form of contrastive ensemble. We isolate the tokens that are more likely under the allocentric framing than under the camera framing. This effectively highlights predictions that genuinely reflect the reference object's perspective rather than the camera's viewpoint. Importantly, since both x_{a} and x_{c} describe the same spatial configuration (from different viewpoints), the visual information remains constant (Figure 2), and the contrastive signal purely captures perspective-dependent reasoning. When questions involve multiple spatial axes, viewpoint verification determines which axis requires bias correction. We apply PCD to options belonging to the affected axis to avoid over-correction. To ensure fair comparison, we apply mean-preserving normalization (Appendix A.1).

4 Experiments

We evaluate our approach on allocentric spatial reasoning benchmarks, comparing against multiple baselines. We demonstrate generalizability using open-source VLMs: Qwen2.5-VL-7B (Bai et al., 2025b) ands InternVL2.5-8B (Chen et al., 2025). We further assess generalization across a broader range of architectures and scales in Sec. 4.2 (Model size and architecture variants).

Model	3DSRBench						OmniSpatial		SpatialMQA		ViewSpatial	
	Overall		front/behind		left/right							
	score	↑	score	↑	score	↑	score	↑	score	↑	score	↑
LLaVA-OneVision	39.25	–	46.51	–	32.09	–	24.63	–	21.85	–	35.75	–
PCD-LLaVA-OneVision	64.21	+24.96	58.43	+11.92	69.91	+37.82	55.67	+31.04	39.29	+17.44	40.38	+4.63
Qwen2.5-VL-3B	38.67	–	42.73	–	34.67	–	44.33	–	20.75	–	30.05	–
PCD-Qwen2.5-VL-3B	58.73	+20.06	49.13	+6.40	68.19	+33.52	60.59	+16.26	37.09	+16.34	42.04	+11.99
Qwen2.5-VL-7B	39.97	–	45.06	–	34.96	–	33.00	–	23.62	–	34.92	–
PCD-Qwen2.5-VL-7B	72.29	+32.32	68.31	+23.25	76.22	+41.26	64.04	+31.04	47.02	+23.40	45.61	+10.69
Qwen2.5-VL-32B	43.29	–	54.65	–	32.09	–	34.98	–	30.24	–	40.97	–
PCD-Qwen2.5-VL-32B	67.10	+23.81	62.79	+8.14	71.35	+39.26	60.10	+25.12	47.02	+16.78	45.49	+4.52
Qwen3-VL-4B	40.12	–	54.36	–	26.07	–	30.05	–	19.65	–	39.07	–
PCD-Qwen3-VL-4B	65.08	+24.96	61.34	+6.98	68.77	+42.70	**66.01**	+35.96	**50.33**	+30.68	44.42	+5.35
Qwen3-VL-8B	44.30	–	58.43	–	30.37	–	33.50	–	24.72	–	39.90	–
PCD-Qwen3-VL-8B	69.12	+24.82	68.02	+9.59	70.20	+39.83	57.64	+24.14	44.59	+19.87	47.98	+8.08
Qwen3-VL-32B	40.98	–	57.27	–	24.93	–	33.00	–	18.10	–	45.25	–
PCD-Qwen3-VL-32B	**76.33**	+35.35	**75.29**	+18.02	**77.36**	+52.43	62.56	+29.56	48.79	+30.69	53.56	+8.31
InternVL2.5-8B	45.17	–	60.17	–	30.37	–	40.89	–	28.48	–	41.92	–
PCD-InternVL2.5-8B	67.82	+22.65	64.53	+4.36	71.06	+40.69	63.55	+22.66	47.46	+18.98	**54.63**	+12.71
InternVL3.5-8B	40.40	–	45.93	–	34.96	–	35.96	–	27.37	–	34.92	–
PCD-InternVL3.5-8B	66.09	+25.69	61.63	+15.70	70.49	+35.53	62.56	+26.60	46.58	+19.21	46.44	+11.52
InternVL3.5-14B	40.40	–	49.42	–	31.52	–	37.44	–	26.27	–	35.04	–
PCD-InternVL3.5-14B	65.80	+25.40	59.88	+10.46	71.63	+40.11	60.10	+22.66	44.15	+17.88	51.54	+16.50
InternVL3.5-38B	39.25	–	51.16	–	27.51	–	30.05	–	24.50	–	42.04	–
PCD-InternVL3.5-38B	74.75	+35.50	72.38	+21.22	77.08	+49.57	61.58	+31.53	48.57	+24.07	54.39	+12.35

Table 2: Performance across different model architectures and scales. Each baseline is paired with its PCD-enhanced counterpart, with ↑ indicating improvement over the baseline. Our method consistently improves performance across all model sizes and architectures.

Configuration	3DSR Bench	Omni Spatial	Spatial MQA	View Spatial
Baseline	39.97	33.00	23.62	34.92
+ PCD only	48.92	12.81	28.70	36.34
+ Full Pipeline	**72.29**	**64.04**	**47.02**	**45.61**

Table 3: Component ablation. Without viewpoint alignment verification, PCD alone shows limited results, demonstrating the necessity of our approach.

Method	Memory	Time
APC-Vis	29629 MiB	17.47s
Ours (Qwen2.5-VL-7B)	18544 MiB	2.02s

Table 4: Memory usage and processing time comparison. Measured on a 1 NVIDIA A100 80GB GPU.

4.1 Experimental Setup

Benchmarks. We evaluate our approach on four real-world image-based spatial reasoning benchmarks. Sample image-question pairs from each benchmark are shown in Figure 3.

3DSRBench (Ma et al., 2025) is a 3D spatial reasoning benchmark constructed for VLMs. It tests allocentric front/behind and left/right relations.

OmniSpatial Benchmark (Jia et al., 2026) was introduced as a benchmark designed to comprehensively evaluate the spatial reasoning capabilities of VLMs. We test questions in the allocentric category of Perspective Taking that query spatial relations.

SpatialMQA (Liu et al., 2025a) is a human-annotated benchmark designed to evaluate the spatial relationship reasoning capabilities. We use the test set and evaluate on allocentric questions. The benchmark includes relative concepts consisting of complex spatial relationship judgments.

ViewSpatial-Bench (Li et al., 2025b) is a comprehensive benchmark designed to evaluate the multi-viewpoint spatial localization capabilities of VLMs. We use the Person Perspective - Relative direction category to assess allocentric spatial relationship judgment abilities. This benchmark consists of complex and detailed spatial relationship questions.

Baselines. We compare against multiple models. **Open-source VLMs:** LLaVA-1.5-7B (Liu et al., 2024a), SpaceLLaVA (Foutter et al., 2024), LLaVA-NeXT (Liu et al., 2024b), LLaVA-OneVision (Li et al., 2025a), Molmo (Deitke et al., 2025), Cambrian-1 (Tong et al., 2024a), InternVL2.5-8B (Chen et al., 2025), and Qwen2.5-VL-7B (Bai et al., 2025b). **Proprietary models:** GPT-5.4 (Singh et al., 2026) and Gemini-2.5-Pro (Comanici et al., 2025). **Grounded VLM:** SpatialRGPT (Cheng et al., 2024). **Fine-tuned**

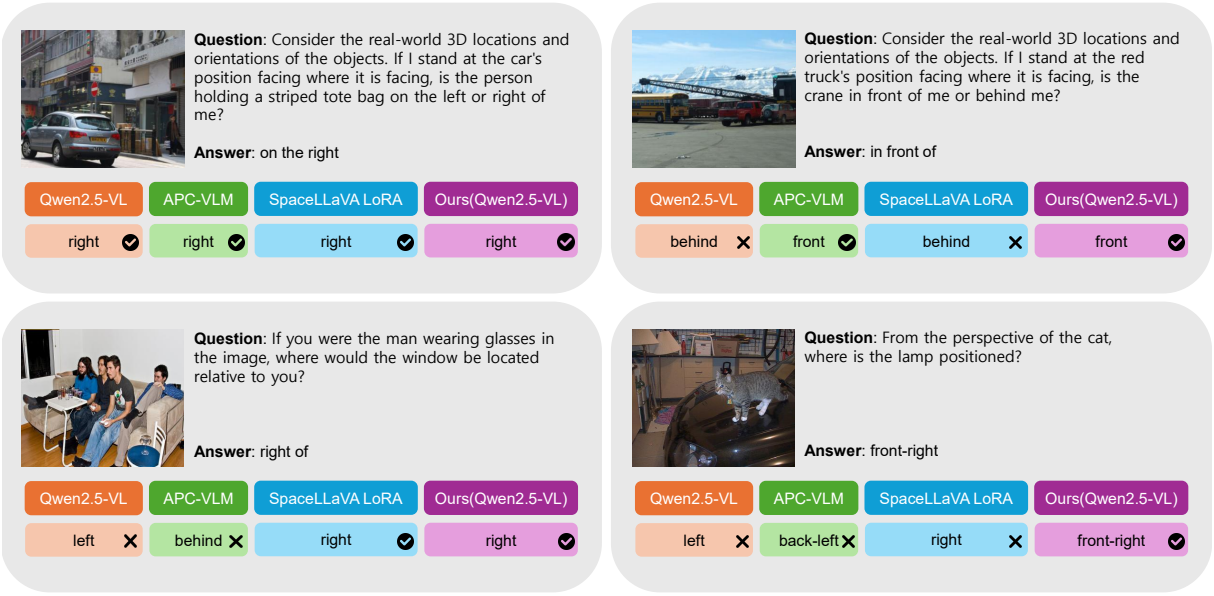


Figure 4: Qualitative comparison on perspective-dependent spatial reasoning. Existing approaches including Qwen2.5-VL (Bai et al., 2025b), APC-Vis (Lee et al., 2025), and fine-tuned SpaceLLaVA (Liu et al., 2025a) fail when reasoning requires viewpoint changes with complex spatial relationships. Our method effectively handles these cases through explicit perspective specification and PCD.

Model: SpaceLLaVA (LoRA fine-tuned with SpatialMQA train dataset). **Training-free framework:** APC-Vis (Lee et al., 2025). Implementation details are in Appendix A.1.

4.2 Results

Main results. Table 1 compares our method against existing models across allocentric spatial relationship reasoning benchmarks. We evaluate our method applied to two representative open-source VLMs: Qwen2.5-VL-7B and InternVL2.5-8B.

As shown in Table 1, most open-source VLMs struggle with the task. This confirms earlier observations (Zhang et al., 2025b; Shiri et al., 2024) that VLMs fail to adopt alternative perspectives. Proprietary models demonstrate relatively lower performance, which can be attributed to camera-centric bias causing errors. Even specialist VLMs designed for spatial reasoning perform poorly. On the other hand, our method achieves state-of-the-art results on 3DSRBench across both subtypes. With Qwen2.5-VL-7B, we attain 76.22% on left/right judgments and 68.31% on front/behind, surpassing the previous best APC-Vis (72.28% and 67.44%).

On OmniSpatial, our approach yields improvements, with Qwen2.5-VL-7B reaching 64.04% and InternVL2.5-8B reaching 63.55%. This substantial improvement demonstrates our method's effectiveness on allocentric perspective-taking tasks.

For SpatialMQA, InternVL2.5-8B achieves 47.46%, competitive with but below the finetuned SpaceLLaVA (51.66%). However, we note that finetuned SpaceLLaVA's performance degrades significantly on other benchmarks (34.10% on 3DSR-Bench L/R vs. our 76.22%), suggesting that fine-tuning alone has limited generalization capability.

On ViewSpatial-Bench, InternVL2.5-8B reaches 54.63%, outperforming all reproducible baselines by substantial margins. Figure 4 shows qualitative comparisons across models on perspective-dependent spatial reasoning questions.

The consistent improvements suggest that our method successfully adapts to different architectures' spatial reasoning characteristics.

Generalization across architectures and scales. We conducted experiments on base open-source VLMs with our methodology applied in Table 2. We experiment with the Qwen2.5-VL family at 3B, 7B (the backbone also used by APC-Vis), and 32B scales, along with InternVL2.5-8B. To further verify generalization to newer and larger backbones, we additionally evaluate on 7 more models spanning 3 families (Qwen3-VL-4B/8B/32B, InternVL3.5-8B/14B/38B, and LLaVA-OneVision), ranging from 4B to 38B parameters. When our proposed methodology was applied to all variants, performance improvements were observed.

Notably, baseline performance does not scale monotonically with parameter count within a family. Within Qwen2.5-VL, the 7B baseline outperforms the 32B baseline on left/right judgments (34.96 vs. 32.09), and this ordering persists after PCD is applied (76.22 vs. 71.35). Qwen3-VL peaks at 8B (44.30 overall) before dropping at 32B (40.98), and InternVL3.5 baselines remain flat to slightly lower than InternVL2.5-8B across scales (39.25–40.40 vs. 45.17). We discuss this pattern further in Appendix C.4.

Component Ablation. Table 3 demonstrates the importance of viewpoint alignment verification. Applying PCD without it shows inconsistent results. While it improves 3DSRBench (+8.95) and SpatialMQA (+5.08), it fails on OmniSpatial (-20.19). Our full pipeline addresses this by viewpoint alignment verification stage. The addition of the stage improves the performance, achieving consistent improvements across all benchmarks. This validates viewpoint verification is essential for robust allocentric reasoning.

To assess the reliability of this gating stage, we measure verification accuracy directly. PCD-Qwen2.5-VL-7B achieves 86.72% (3DSRBench) and 90.42% (OmniSpatial) alignment classification accuracy. Not all misclassifications propagate to the final answer as some coincide with cases where camera-frame and reference-frame answers agree. The actual downstream impact is 10.97%/8.87% (Appendix C.2). This indicates the verification step is a reasonably reliable, far-from-random gate, explaining the consistent end-to-end gains.

We use our PCD formulation (Eq. 2) rather than the generalized formula, and employ a single model for both perspectives. Ablation studies (Table 6 in Appendix A.1) confirm these choices achieve better performance with lower memory overhead.

We additionally examine PCD's behavior under oblique (non-flip) viewpoint changes and on the out-of-domain MindCube benchmark. See Appendices C.3 and C.5.

Efficiency. We compared memory usage and inference time with APC-Vis (Lee et al., 2025), another training-free methodology in Table 4. The experiments were conducted on 3DSRBench. Since APC-Vis uses multiple external vision modules, our methodology achieves a 37.41% reduction in memory compared to APC-Vis, even when using the same backbone VLM, Qwen2.5-VL-7B-Instruct. Furthermore, in terms of inference time, APC-Vis requires 17.47 seconds while our methodology takes 2.02 seconds per sample.

5 Conclusion

In this work, we address the persistent challenge of allocentric spatial reasoning in VLMs, by identifying camera viewpoint bias as the predominant cause of performance gaps. While fine-tuning demonstrates poor generalization to unseen benchmarks, training-free frameworks employing multiple external models suffer from high memory overhead and slow inference.

We propose a pipeline with Perspective-conditional Prompt-based Contrastive Decoding (PCD), a training-free method that explicitly suppresses the egocentric bias in VLMs. By factoring out bias, PCD enables models to leverage their spatial reasoning capabilities for allocentric perspective-taking without additional training or external modules. Our approach can operate with a single VLM, maintaining low memory footprint and fast inference. Extensive experiments on allocentric spatial reasoning benchmarks demonstrate that PCD outperforms most of the baselines. Our findings suggest that camera viewpoint bias in allocentric spatial relationship reasoning, a defined and practically important class of single- and few-image tasks in robotics and embodied AI, can be mitigated through lightweight, training-free strategies.

Limitations

We position PCD as a training-free technique for allocentric spatial relationship reasoning over single or few-image inputs, rather than a general-purpose spatial reasoning solution. As a multi-stage pipeline built on the base VLM's own spatial perception, PCD's performance is bounded by the backbone's spatial understanding, and errors at any stage can propagate to the final answer. Since the verification stage is the driver of both overall performance and remaining errors, improving gating-stage robustness, e.g., via ensemble-based verification, is an important direction for future work. PCD may also face additional challenges on extremely complex scenes with dense object configurations or severe occlusions.

Our method focuses on relative spatial relationships rather than quantitative properties such as precise distances or metric coordinates, and does not address video or multi-hop spatial inference.

We leave these extensions to future work.

PCD does not compute an explicit geometric transform between the camera and reference viewpoints. It instead suppresses the model's camera-aligned default answer. A fully geometry-aware correction for intermediate orientations remains an open direction.

Finally, although performance is relatively stable across a range of hyperparameter α, the optimal value varies by model size and spatial axis, so some tuning is needed to reach peak performance on a new benchmark or backbone.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25520490).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. 2025. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pages 75602–75637.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

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. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104.

Matthew Foutter, Daniele Gammelli, Justin Kruger, Ethan Foss, Praneet Bhoj, Tommaso Guffanti, Simone D'Amico, and Marco Pavone. 2024. Spacellava: a vision-language model adapted to extraterrestrial applications. *arXiv preprint arXiv:2408.05924*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Google. 2026. Gemini 3.1 flash image (nano banana 2). Google AI Studio.

Gracjan Góral, Alicja Ziarko, Michal Nauman, and Maciej Wolczyk. Seeing through their eyes: Evaluating visual perspective taking in vision language models. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. 2026. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. 2025. Is 'right'right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14257–14267. IEEE.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. 2025. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9241–9251.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, and 1 others. 2025b. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. 2026. Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 40096–40107.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025a. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025b. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *arXiv preprint arXiv:2511.15722*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024d. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. 2026. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *Advances in Neural Information Processing Systems*, 38:140751–140774.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *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.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. 2025. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4567–4576.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. 2026. Mindcube: Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. 2025b. Site: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9058–9069.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. 2024. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv preprint arXiv:2412.18605*.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. 2025a. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11591–11609.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025b. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.

A PCD: Perspective-Conditional Prompt-based Contrastive Decoding

A.1 Implementation Details

Model. We implement PCD on open-source VLMs spanning five architecture families: Qwen2.5-VL (3B/7B/32B, Bai et al. (2025b)), Qwen3-VL (4B/8B/32B, Bai et al. (2025a)), InternVL2.5-8B (Chen et al., 2025), InternVL3.5 (8B/14B/38B, Wang et al. (2025a)), and LLaVA-OneVision (Li et al., 2025a). Our main results (Table 1) use Qwen2.5-VL-7B-Instruct and InternVL2.5-8B. The remaining models are used for the scale and architecture generalization study (Table 2, Section 4.2). All models are open-source, making our approach accessible for research. We set do_sample=False for all experiments. All experiments were conducted on 1-2 NVIDIA A100 80GB GPUs. Our code is available at https://github.com/hjchung1/PCD.

Hyperparameter. To evaluate the robustness of our approach, we conducted an ablation study on the hyperparameter α using the 3DSRBench. All experiments were performed using Qwen2.5-VL as the base model to ensure controlled comparison. We varied α across 0.5, 1.0, 2.0, 4.0.

Table 5 presents the results of hyperparameter sweep. Our method demonstrates consistent improvements over the baseline across all tested α, with performance gains ranging from 4.48 to 32.32. The optimal value of $\alpha = 4.0$ for our method based on Qwen2.5-VL-7B-Instruct achieves the best performance. Importantly, performance remains relatively stable across $\alpha \in [1.0, 4.0]$. This robustness validates that our improvements reflect genuine methodological advances.

As shown in Table 5, different spatial axes may benefit from different α values. For our main results (Table 1), we select axis-specific values within this robust range $[1.0, 4.0]$. Across benchmarks, left/right relations consistently use higher values ($\alpha \approx 2{-}4$) while front/behind and above/below use lower values ($\alpha \approx 1{-}2$), reflecting the higher perspective-dependence of horizontal spatial reasoning. Complete hyperparameter specifications are provided in our code repository.

For the additional architectures and scales evaluated in Table 2 (Qwen3-VL, InternVL3.5, and LLaVA-OneVision), in 3DSRBench we use $\alpha = 1$. For OmniSpatial and SpatialMQA we reuse each model's corresponding original-family α, for ViewSpatial-Bench we reuse the α tuned for InternVL2.5-8B. As our sensitivity analysis (Table 5) shows performance is relatively stable across $\alpha \in [1.0, 4.0]$, we expect these untuned values to yield conservative estimates of PCD's true gains on these backbones. Model-specific tuning would plausibly improve results further.

Model	α	Overall	Front/ Behind	Left/ Right	Δ
3B	Base	38.67	42.73	34.67	–
	0.5	43.15	45.35	40.97	+4.48
	1.0	56.13	47.67	64.47	+17.46
	2.0	57.14	47.67	66.48	+18.47
	4.0	**58.73**	**49.13**	**68.19**	**+20.06**
7B	Base	39.97	45.06	34.96	–
	0.5	52.24	66.28	38.40	+12.27
	1.0	70.85	**69.19**	72.49	+30.88
	2.0	71.57	68.02	75.07	+31.60
	4.0	**72.29**	68.31	**76.22**	**+32.32**
32B	Base	43.29	54.65	32.09	–
	0.5	55.27	63.95	46.70	+11.98
	1.0	66.67	**64.24**	69.05	+23.38
	2.0	**67.10**	62.79	**71.35**	**+23.81**
	4.0	66.96	62.50	**71.35**	+23.67

Table 5: Hyperparameter sensitivity analysis on 3DSR-Bench. We evaluate the impact of α across different Qwen2.5-VL model sizes. Our method consistently outperforms the baseline across all configurations, with improvements ranging from +4.48 to +32.32 points. The optimal α varies by model size, demonstrating robust performance across the hyperparameter space.

Design choice ablation We conduct ablation studies to validate our design choices for PCD on 3DSRBench. Results are shown in Table 6.

We compare our PCD formulation (Eq. 2) with a generalized CD formula. Our variant achieves 72.29% overall accuracy, outperforming the generalized formula (67.10%) by 5.19 percentage points. The improvement is particularly significant on the left/right axis (76.22% vs. 65.90%), where camera perspective bias is most pronounced.

We also explore whether using different-sized models for allocentric and camera perspectives can improve performance. Specifically, we test Qwen2.5-VL-7B&3B, Qwen2.5-VL-32B&7B which use larger model for allocentric prompt and smaller model for camera prompt. Both multi-model configurations show improvements over the baseline (39.97%) but underperform our single-model approach (72.29%).

This indicates that explicit prompt-based perspective contrast within a single model is more

Configuration	Overall	front/ behind	left/ right
Baseline (Qwen2.5-VL-7B)	39.97	45.06	34.96
Effect of formulation:			
Generalized formula	67.10	68.39	65.90
Our variant	**72.29**	**68.31**	**76.22**
Multi-model variants:			
Qwen2.5-VL-7B&3B	67.82	67.73	67.91
Qwen2.5-VL-32B&7B	66.67	64.24	69.05

Table 6: Design choice ablations on 3DSRBench. Our formulation outperforms the generalized CD formula, and single-model approach is both more efficient and effective than multi-model variants.

effective than relying on model size differences. Moreover, our single-model formulation offers significant practical advantages, including memory efficiency, computational efficiency, and simplicity, while also achieving superior accuracy compared to the multi-model alternatives.

Numerical Stability via Log-Space Formulation Following standard practice in neural text generation (Holtzman et al., 2019; Li et al., 2023), we implement PCD in log-probability space for numerical stability:

$$s_{\text{PCD}}(y) = \log p_\theta(y \mid v, x_{\text{a}}) - \alpha \cdot \log p_\theta(y \mid v, x_{\text{c}}) \tag{3}$$

This is ranking-equivalent to Equation 2, since $\arg\max$ is invariant to monotonic transformations and additive constants:

$$\arg\max_y p_{\text{PCD}}(y) = \arg\max_y \left[\log p_\theta(y \mid v, x_{\text{a}}) - \alpha \cdot \log p_\theta(y \mid v, x_{\text{c}})\right] \tag{4}$$

Extension to Multi-Axis Questions Some questions involve multiple spatial axes simultaneously. Viewpoint verification may indicate that only certain axes are affected by bias. When PCD is applied selectively to the affected axes, options requiring PCD ($\mathcal{O}_{\text{PCD}}$) undergo contrastive decoding while unaffected options ($\mathcal{O}_{\text{orig}}$) use original scores. The contrastive term shifts the magnitude of scores for $\mathcal{O}_{\text{PCD}}$, creating unfair comparison across axes.

We apply mean-preserving normalization to prevent it. Let:

$$s(y) = \begin{cases} \log p_\theta(y \mid v, x_{\text{a}}) - \alpha \cdot \log p_\theta(y \mid v, x_{\text{c}}) & y \in \mathcal{O}_{\text{PCD}} \\ \log p_\theta(y \mid v, x_{\text{a}}) & y \in \mathcal{O}_{\text{orig}} \end{cases} \tag{5}$$

Method	Accuracy
Baseline	34.92
All PCD	36.34
Selective PCD w/o mag. norm	40.86
Selective PCD w/ mag. norm	45.61

Table 7: Ablation study on magnitude normalization for multi-axis questions of ViewSpatial. Selective PCD applies CD only to axes affected by viewpoint bias. Magnitude normalization prevents systematic score shifts between axis groups.

We compute the mean shift over options requiring PCD:

$$\Delta = \frac{1}{|\mathcal{O}_{\text{PCD}}|} \sum_{y \in \mathcal{O}_{\text{PCD}}} \left[\log p_\theta(y \mid v, x_{\text{a}}) - s(y)\right] \tag{6}$$

and apply $s_{\text{final}}(y) = s(y) + \Delta$ for $y \in \mathcal{O}_{\text{PCD}}$. This ensures both axis groups have comparable score magnitudes while preserving the relative ranking induced by contrastive decoding. Table 7 demonstrates that this normalization is essential.

B Pipeline

B.1 Evaluation Details

Our work focuses on improving VLM performance on allocentric spatial reasoning tasks, where the model must determine directional relationships between objects from a specified viewpoint. We evaluate on four benchmarks that align with this scope.

On 3DSRBench (Ma et al., 2025), following the original benchmark setup, we employ CircularEval (Liu et al., 2024d), which assesses response consistency by permuting answer options. A VLM is considered correct only if it selects the right answer across all permutations. We evaluate on the `orientation_on_the_left` and `orientation_in_front_of` categories, both of which test allocentric directional reasoning.

For ViewSpatial-Bench (Li et al., 2025b), we evaluate on the Person Perspective—relative direction category, which requires allocentric directional judgments including fine-grained eight-directional relations.

Omnispatial Benchmark (Jia et al., 2026) includes diverse spatial reasoning tasks with mixed question types. We evaluate on Perspective Taking questions with allocentric directional queries.

For SpatialMQA (Liu et al., 2025a), we evaluate on allocentric questions from the test split where the target object is visually present in the image.

B.2 Details on Text Prompts

In this section, we present the text prompts used in our pipeline. First, in the object extraction stage (§3.1), we provide exemplar question-answer pairs as in-context demonstrations to steer the VLM toward generating responses in the target format.

We used the following prompt template:

> Extract the "reference object" and "target object" from a spatial reasoning question.
>
> - reference object: the object whose position and facing direction defines the viewpoint
>
> - target object: the object whose position is being asked about
>
> Examples:
>
> Q: If you were a woman walking on the beach, on which side of you would the sunbed be? A: reference object: woman walking on the beach, target object: sunbed
>
> Q: If you are the athlete in the image, where is the cat located relative to you? A: reference object: athlete, target object: cat
>
> Q:

Next, in the viewpoint alignment verification stage (§3.1), we utilize the VLM's visual perception to confirm alignment between the reference object's facing direction and the camera viewpoint. The prompt used for this stage is as follows:

> Question: In the image, are you looking at the `{reference_object}`'s back side?

C Additional Analysis

C.1 Analysis on Training-free Baselines

We analyze two training-free approaches that could potentially address allocentric spatial reasoning: multi-model pipelines and VCD.

C.1.1 APC-Vis: Multi-model Framework

In this section, we further discuss the training-free framework baseline introduced in Section 4.1. While our method leverages the distinction between allocentric and egocentric formulations, an intuitive approach is to uniformly convert all spatial queries to the egocentric perspective. APC-Vis (Lee et al., 2025) adopts this strategy by transforming allocentric questions into egocentric form and generating simplified visual depictions from the egocentric viewpoint, which are then provided to the VLM.

This framework is built on Qwen2.5-VL-7B-Instruct as the backbone VLM and relies on multiple external foundation models, including GroundingDINO (Liu et al., 2024c), SAM (Kirillov et al., 2023), DepthPro (Bochkovskiy et al., 2025), and OrientAnything (Wang et al., 2024). While APC-Vis demonstrates the potential of visual abstraction for allocentric spatial reasoning, its pipeline chains multiple independently-trained external modules.

As illustrated in Figure 5, failures arise from various sources such as inaccurate object abstraction, misaligned visual prompt generation, or inconsistencies between the generated visual prompts and the VLM's final predictions. These error cases reveal a trade-off in multi-component pipelines. While modular design enables sophisticated reasoning workflows, performance becomes inherently bounded by the weakest component in the chain. Each additional module introduces potential failure points that compound through the pipeline.

This observation motivates our design choice to minimize external dependencies while achieving competitive performance through strategic prompt reformulation and viewpoint verification. This does not mean PCD is immune to error propagation. As a multi-stage pipeline, an early misclassification can still propagate to the final answer (Appendix C.2). The distinction we draw is narrower. APC-Vis's external modules share no representation with each other or with the backbone VLM, so cross-module disagreement adds error on top of any single module's mistakes, which is visible in Figure 5, where failures often trace to inconsistencies between modules. PCD's stages instead operate within a single backbone VLM, and its viewpoint-verification stage alone already reaches 86.72%/90.42% accuracy on 3DSR-Bench/OmniSpatial (Section 4.2), avoiding this cross-module mismatch even though sequential dependency remains. Reducing this residual error further, e.g. via ensemble-based verification, is a direction for future work.

C.1.2 VCD: Image-space Perturbation

Visual Contrastive Decoding (VCD) (Leng et al., 2024) is another training-free method that uses contrastive decoding. However, VCD operates in vi-

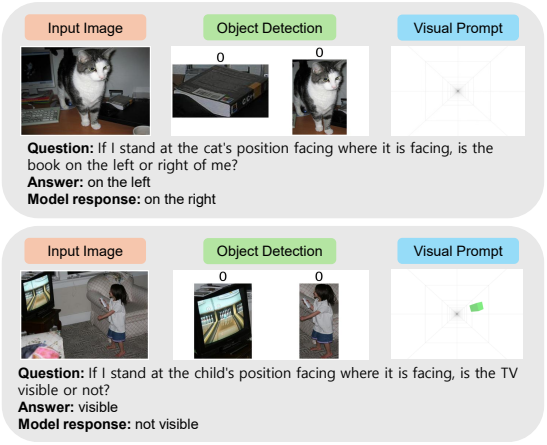


Figure 5: Representative failure cases of APC-Vis. The first example demonstrates failure due to inaccurate visual prompt generation. The second example shows VLM prediction error despite a reasonably generated visual prompt, indicating limitations in the model's spatial reasoning capability.

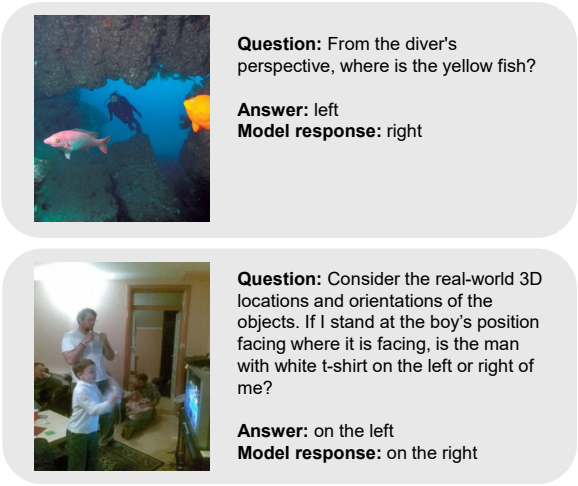


Figure 6: Qualitative examples of the two recurring error patterns in Table 9. Top: a *viewpoint verification* error, where camera-object alignment is misjudged. Bottom: a *PCD over-suppression* case, where the camera-frame answer was in fact correct.

sual space (contrasting clean vs. noisy images) to address object hallucinations, while our PCD operates in semantic space (contrasting allocentric vs. egocentric framings) to address perspective bias. To verify that visual perturbation is poorly suited for allocentric reasoning, we evaluate VCD on two officially supported base models.

We evaluate VCD using the official implementation with cd_alpha=1.0, cd_beta=0.1, and noise_step=500 on LLaVA-1.5-7B and QwenVL to verify that visual perturbation cannot address allocentric reasoning challenges.

Results. As shown in Table 8, our PCD achieves 76.22% on left/right judgments, a 54.44-point improvement over VCD. This substantial gap supports our hypothesis: allocentric reasoning requires semantic-level intervention that directly targets perspective bias, rather than visual-level perturbation.

Analysis. VCD shows inconsistent behavior. While it improves LLaVA's extremely weak left/right performance (5.44% $\rightarrow$ 21.78%), it degrades ViewSpatial (-11.04 points), and all configurations remain far below random chance (21.78% vs. 50% on L/R, 29.65% vs. 50% on F/B). This mixed pattern reveals two key limitations. Camera viewpoint bias is predominantly an interpretation issue rather than a perceptual one. While VLMs typically perceive spatial configurations accurately, they frequently reason from the camera's reference frame instead of the specified object's perspective. Gaussian noise does not alter the reference frame. It only degrades visual clarity. This cannot systematically correct perspective bias. Second, spatial reasoning requires fine-grained visual cues such as object orientation, body pose, and relative positions. Visual noise degrades exactly these cues, making accurate spatial judgment more difficult rather than correcting perspective bias.

Summary. These results confirm that both multi-model pipelines (APC-Vis) and visual perturbation methods (VCD) face limitations for allocentric reasoning. PCD combines the efficiency advantages of single-model inference with direct targeting of camera viewpoint bias.

C.2 Error Pattern Analysis

While our method demonstrates substantial improvements over baselines on allocentric spatial reasoning tasks, we conduct a qualitative error decomposition to identify remaining challenges and opportunities for future work.

Error decomposition. We decompose errors on 3DSRBench and OmniSpatial (PCD-Qwen2.5-VL-7B) into five mutually exclusive categories, grouped into two stages. The *gating stage*: (1) *object extraction* errors, where the wrong reference or target object is identified, and (2) *viewpoint verification* errors, where alignment is misjudged. The *post-verification stage*: (3) *direct generation* errors, where verification correctly judges alignment but the direct answer is wrong. (4) *PCD over-suppression*, where verification correctly judges

Approach	3DSRBench		Omni Spatial	Spatial MQA	View Spatial
	F/B	L/R			
Baseline (for reference)					
LLaVA-1.5-7B	28.78	5.44	25.62	25.83	39.31
Visual Contrastive Methods					
VCD (LLaVA)	29.65	21.78	25.12	24.50	28.27
VCD (QwenVL)	27.33	25.79	25.62	24.94	32.19
Random	50.00	50.00	25.00	25.59	25.00
Our Semantic Contrastive Method					
PCD (Qwen2.5-VL)	**68.31**	**76.22**	**64.04**	**47.02**	**45.61**
Δ vs VCD (best)	+38.66	+50.44	+38.42	+22.08	+13.42

Table 8: Comparison of visual contrastive (VCD) vs. semantic contrastive (PCD) approaches on allocentric spatial reasoning. VCD improves over the weak LLaVA baseline on some metrics but remains at or below random chance on directional judgments across both base models. Our PCD achieves improvements over VCD, validating that allocentric reasoning requires semantic-level intervention that directly targets perspective bias.

Error type	3DSRBench	OmniSpatial
Correct	74.31	64.04
Object extraction	0	10.34
Viewpoint verification	10.97	8.87
Direct generation	6.71	4.93
PCD over-suppression	6.42	5.91
Residual decoding	1.59	5.91

Table 9: Error decomposition on 3DSRBench and OmniSpatial (PCD-Qwen2.5-VL-7B).

	3DSRBench	OmniSpatial
Qwen2.5-VL-7B	38.76	31.81
+ PCD	**75.25**	**61.36**

Table 10: Accuracy on non-flip subsets. PCD's gains persist even when the camera-to-reference mapping is a rotation rather than a flip.

misalignment but the camera-frame answer was in fact correct, and (5) *residual decoding* errors, where PCD is applied correctly but the final answer is still wrong. Table 9 reports the results.

The dominant single error category differs between benchmarks. Viewpoint verification is the largest source on 3DSRBench (10.97%), which uses fixed-template questions, while object extraction dominates on OmniSpatial (10.34%), whose free-form phrasing makes few-shot extraction harder. We selected these two benchmarks because they represent contrasting question formats. The fact that the dominant error source shifts accordingly suggests that this gating-stage bottleneck pattern generalizes across question formats rather than being an artifact of a single benchmark.

Figure 6 illustrates one example from each stage. The top case is a viewpoint verification failure. The pipeline's gating decision itself is wrong, so improving the gating stage (e.g., through ensemble-based verification instead of a single prompt) would directly address it. The bottom case is PCD over-suppression. The gate correctly detects misalignment, but contrastive decoding overcorrects an answer that the camera frame already got right. This suggests that certain spatial configurations are more naturally expressed from the camera viewpoint than others, pointing toward adaptive strategies that weigh evidence from both framings rather than always trusting the allocentric one once misalignment is detected.

These errors contribute non-trivially to the remaining failures, and their relative importance is benchmark-dependent, motivating improvements on both fronts.

C.3 Analysis on Viewpoint Configurations

We additionally verify PCD's robustness on non-flip configurations. Suppressing the camera-aligned answer cleanly recovers the correct one in flip case. When the reference object faces sideways or obliquely, the camera-to-object mapping is a rotation across axes rather than a flip, and it is not immediately obvious that suppressing the camera's answer still points to the correct one.

We test this by manually inspecting subsets of 3DSRBench and OmniSpatial where the reference object is sideways or oblique to the camera, excluding flip cases entirely.

As shown in Table 10, PCD's gains persist on this subset, improving accuracy from 38.76 to

75.25 on 3DSRBench and from 31.81 to 61.36 on OmniSpatial. This indicates that PCD's benefit is not confined to the flip case. As additional evidence, ViewSpatial-Bench's 8-way ground truth (front, front-left, left, back-left, back, back-right, right, front-right) includes diagonal labels that are only correct under non-collinear geometry (i.e., rotation rather than flip), and PCD improves accuracy on this benchmark as well (Qwen2.5-VL: 34.92→45.61, InternVL2.5: 41.92→54.63).

We emphasize that PCD performs no explicit geometric correction, nor do we claim it does. It suppresses the model's camera-aligned bias directly (Sec. 3.4), which is why its benefit generalizes past the flip case empirically rather than by geometric construction. Extending PCD toward fully geometry-aware correction for intermediate orientations remains an open direction for future work.

C.4 Non-monotonic Scaling of Allocentric Reasoning

Table 2 shows that PCD-enhanced accuracy does not increase monotonically with parameter count within a model family (e.g., PCD-Qwen2.5-VL-7B outperforms PCD-Qwen2.5-VL-32B on 3DSR-Bench overall: 72.29 vs. 67.10). We verify this is a property of the underlying VLMs' allocentric reasoning rather than an artifact introduced by PCD. The same ordering is already present in the pre-PCD baselines (Qwen2.5-VL-7B baseline 39.97 vs. 32B baseline 43.29 on 3DSRBench overall, but 34.96 vs. 32.09 on left/right specifically). The pattern recurs across the additional model families in Table 2. Mid-sized variants (Qwen3-VL-8B, InternVL3.5-8B/14B) match or exceed their largest counterparts (Qwen3-VL-32B, InternVL3.5-38B) on baseline 3DSRBench accuracy.

Since PCD's gain Δ is mechanically tied to each model's own camera-viewpoint bias magnitude, its relationship with scale inherits this non-monotonicity rather than following one fixed trend. Within Qwen2.5-VL, Δ peaks at 7B on 3 of 4 benchmarks, whereas for Qwen3-VL, Δ increases monotonically on ViewSpatial-Bench. A systematic correlation analysis across all 11 models and 8 benchmark$\times$factor combinations, with multiple-comparison correction, finds only one robust relationship. Baseline accuracy is negatively correlated with OmniSpatial gain ($r = -.855$, $p = .001$, $q < .05$). An apparent positive correlation between scale and 3DSRBench gain ($r = +.634$, uncorrected $p = .036$) does not survive correction and should be read as suggestive only.

We interpret this non-monotonicity as evidence that camera viewpoint bias is primarily shaped by training-data composition and instruction-tuning distribution rather than raw model capacity. Fully characterizing the interaction between scale, model family, and allocentric reasoning ability is left to future work.

Method	Accuracy
Qwen2.5-VL-7B	34.45
+ All-PCD, x_c: reference-visible frame	36.92
+ All-PCD, x_c: reference's given side frame	36.17

Table 11: Pilot results on MindCube Other's Perspective Taking-2, $\alpha = 1.0$, no Stage 2 verification.

C.5 Pilot Study: Scope Boundary on Multi-view Spatial Reasoning

We conduct a small-scale pilot on MindCube (Wang et al., 2026), one of the benchmarks we identified as falling outside PCD's intended scope. We emphasize that this is a scope-boundary probe rather than a claim of generalization. MindCube questions are answered from multiple simultaneous images, which breaks two premises PCD relies on (Sec. 3.1), a single-image-verifiable reference viewpoint (Stage 2), and a well-defined single camera-perspective prompt x_c to contrast against.

Setup. We evaluate on MindCube's *Other's Perspective Taking-2* subtask ("Level-2"), which requires mental rotation to determine how a spatial relationship between two objects transforms under a shift to another agent's viewpoint (Wang et al., 2026), as it is the subtask structurally closest to our formulation. We use Qwen2.5-VL-7B-Instruct, unchanged from our main experiments.

Because Stage 2 of our pipeline presupposes a single image in which the camera-object alignment can be visually verified, it is not well-defined when the input is a multi-image observation, so we omit it and apply PCD to all instances, with $\alpha = 1.0$. Also, since MindCube provides multiple images per question, the camera-perspective prompt x_c is not uniquely defined. We therefore test two natural constructions: (i) the frame in which the reference object is visible, and (ii) a frame showing the reference object's certain side.

Results. Both constructions improve modestly over the baseline (+2.47 and +1.72 points, Table 11), in contrast to the +20 to +40 point gains

PCD achieves on single-image benchmarks (Table 1). The two camera-prompt definitions yield nearly identical accuracy, suggesting the small residual gain is not sensitive to how x_c is constructed, but rather reflects a ceiling imposed by the setting itself.

This pattern is consistent with, rather than a counterexample to, the scope boundary we describe. PCD's correction is mechanically tied to an accuracy asymmetry favoring the camera-aligned (egocentric) response over the allocentric one (Sec. 3.1). MindCube's own reported results suggest this asymmetry does not hold in this setting. For Qwen2.5-VL-7B-Instruct, self-perspective (egocentric) accuracy is *lower* than the level-2 other-perspective (allocentric) accuracy (28.77 vs. 36.81, as reported in Wang et al. (2026)). Subtracting a camera-aligned signal that is not the dominant bias source in this setting yields only a small, formulation-insensitive gain, as observed here, rather than the substantial correction PCD provides when the asymmetry does hold.

We view this as preliminary evidence that extending PCD to multi-view and video settings requires more than adapting Stage 2 to a multi-image input format. It first requires establishing whether, and in which direction, a camera-aligned bias asymmetry exists in such settings, since PCD's contrastive mechanism presupposes this asymmetry rather than diagnosing it. A viable path forward is likely a cross-frame, scene-memory-based verification mechanism that can establish this asymmetry directly from multi-image input, rather than a direct application of the current single-image pipeline. We leave this to future work.