PREDICTING CAMERA POSE FROM PERSPECTIVE DESCRIPTIONS FOR SPATIAL REASONING

Xuejun Zhang, Aditi Tiwari, Zhenhailong Wang & Heng Ji
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
{xuejunz2, aditit5, wangz3, hengji}@illinois.edu

ABSTRACT

Multi-image spatial reasoning remains challenging for current multimodal large language models (MLLMs). While single-view perception is inherently 2D, reasoning over multiple views requires building a coherent scene understanding across viewpoints. In particular, we study perspective taking, where a model must build a coherent 3D understanding from multi-view observations and use it to reason from a new, language-specified viewpoint. We introduce CAMCUE, a pose-aware multi-image framework that uses camera pose as an explicit geometric anchor for cross-view fusion and novel-view reasoning. CAMCUE injects per-view pose into visual tokens, grounds natural-language viewpoint descriptions to a target camera pose, and synthesizes a pose-conditioned imagined target view to support answering. To support this setting, we curate CAMCUE-DATA with 27,668 training and 508 test instances pairing multi-view images and poses with diverse target-viewpoint descriptions and perspective-shift questions. We also include human annotated viewpoint descriptions in the test split to evaluate generalization to human language. CAMCUE improves overall accuracy by 9.06% and predicts target poses from natural-language viewpoint descriptions with over 90% rotation accuracy within 20° and translation accuracy within a 0.5 error threshold. This direct grounding avoids expensive test-time search-and-match, reducing inference time from 256.6s to 1.45s per example and enabling fast, interactive use in real-world scenarios.

1 INTRODUCTION

Spatial intelligence moves beyond single-image perception and naive multi-image aggregation. Rather than treating each view as an independent 2D snapshot, an agent needs to connect views via their spatial relationships to form a coherent 3D understanding that supports reasoning beyond the observed images (Chen et al., 2024a; Gholami et al., 2025; Wang et al., 2025; Yin et al., 2025; Zhao et al., 2025; Lee et al., 2025; Yeh et al., 2025; Yang et al., 2025b). Humans do this naturally: when told "sit on the sofa behind the black table," we can mentally relocate to that viewpoint and imagine what we would see, then answer questions from that perspective (Wang, 2012; Meilinger et al., 2011), which is illustrated by Figure 1. However, current multimodal large language models (MLLMs) still struggle with this kind of perspective taking. Even with multiple context images, they often fail to reliably ground a language-specified viewpoint and reason from the intended perspective (Lee et al., 2025; Yeh et al., 2025; Xu et al., 2025; Yin et al., 2025). This gap motivates our study of language-guided viewpoint grounding for multi-view spatial reasoning. We study perspective-shift reasoning where the target viewpoint is specified in natural language. Given multiple context images and a question, the model needs to ground the description to a target camera pose and answer from that perspective.

A recent line of work tackles perspective-shift reasoning by augmenting MLLMs with generative world models that actively synthesize additional observations at inference time (Lee et al., 2025; Yang et al., 2025c). While promising, existing pipelines are often built around a single reference view and do not effectively integrate multiple contextual images as a unified source of evidence (Lee et al., 2025; Yang et al., 2025c). In addition, most controllable generators are largely query-agnostic,

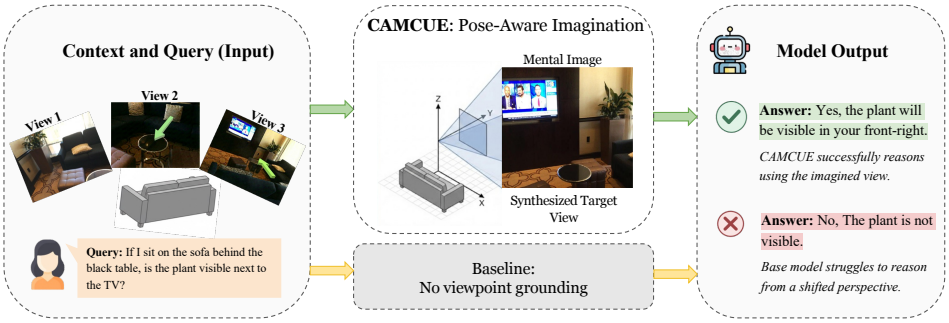


Figure 1: **Perspective-shift reasoning with CamCue.** Given multi-view context images, CamCue maps a natural-language viewpoint description to an explicit target camera pose and synthesizes the corresponding target view for reliable spatial reasoning.

which can produce imagined views that are irrelevant or even inconsistent with the downstream question (Yang et al., 2025c). Many of these methods rely on expensive test-time procedures such as iterative search or multiple candidate rollouts to obtain a useful imagined view, resulting in high latency and limited practicality. Finally, off-the-shelf novel-view synthesis is typically pose-conditioned, whereas MLLMs do not reliably infer target camera poses from natural language description, leaving a mismatch between language-driven viewpoint specification and pose-controlled generation (Jin et al., 2024; Zhou et al., 2025).

To address these limitations, we introduce CamCue, a pose-aware multi-image MLLM framework that can predict the camera pose of the language-specified target perspective. Camera pose provides a compact, explicit representation of viewpoint that situates each image in a shared 3D coordinate frame, making inter-view geometry directly usable for multimodal reasoning (Liao et al., 2025; Zhao et al., 2025). Our key design principle is to make viewpoint an explicit geometric anchor for multi-view reasoning. We start by injecting per-view camera information into the corresponding visual features, so that the model can align evidence across images through geometry rather than treating each image as individual input. We then interpret the natural-language target-perspective description by mapping it to a concrete target camera pose, which specifies where the model should "mentally stand" to answer the question. Conditioned on this predicted pose, we further synthesize the corresponding target-view image and treat the imagined observation as additional evidence for answering. This tight coupling between language-specified perspective, pose prediction, and pose-conditioned view synthesis strengthens multi-image fusion and substantially improves performance on perspective-shift spatial reasoning.

To support this setting, we curate CamCue-Data, a dataset tailored to perspective-shift reasoning. CamCue-Data contains 27,668 training instances and 508 test instances, and pairs multi-view images and per-view camera poses with diverse natural-language target-perspective descriptions, including human-annotated descriptions, and questions that require answering from the specified viewpoint. On this benchmark, CamCue yields substantial gains on perspective-shift spatial reasoning, improving overall accuracy by 9.06%. It also predicts target camera poses directly from natural-language descriptions with strong accuracy, achieving over 90% rotation accuracy within 20° and translation accuracy at t@0.5. Moreover, by predicting an explicit target pose, CamCue avoids expensive test-time search-and-match used by prior methods, reducing inference time from 256.6 seconds to 1.45 seconds per example and enabling fast, interactive use in real-world scenarios. Beyond CamCue-Data, CamCue also improves performance on general multi-image spatial reasoning benchmarks such as MindCube Tiny (Yin et al., 2025) and MMSI (Yang et al., 2025b).

In summary, our contributions are listed as follows:

- We propose CamCue, a pose-aware multi-image MLLM framework that injects per-view camera information into the corresponding visual features, enabling geometry-aware fusion across views for spatial reasoning.
- CamCue can map a natural-language target-view description to an explicit target camera pose, providing a concrete viewpoint representation for answering from the specified perspective.

- Conditioned on the predicted target pose, CAMCUE synthesizes the corresponding target-view image and feeds it back as additional evidence, substantially improving perspective-shift reasoning ability.
- We curate CAMCUE-DATA, a dataset tailored to perspective-shift reasoning that pairs multi-view images with diverse, detailed natural-language camera viewpoint descriptions, including human-annotated descriptions, and target-view questions that require reasoning from the described viewpoint.

2 RELATED WORK

2.1 MULTI-IMAGE SPATIAL REASONING BENCHMARKS

Multi-image spatial reasoning is a key probe for evaluating spatial intelligence in MLLMs, as it requires integrating partial observations from multiple viewpoints into a coherent and viewpoint-consistent scene understanding. Recent benchmarks reveal substantial gaps between current MLLMs and human performance, with models often struggling to fuse evidence across views and maintain consistent spatial beliefs. Representative datasets include MindCube (Yin et al., 2025), SpatialBench (Xu et al., 2025), MMSI-Bench (Yang et al., 2025b), All-Angles Bench (Yeh et al., 2025), and ViewSpatial-Bench (Li et al., 2025). Surveys and diagnostic studies further organize these benchmarks by cognitive demands and emphasize that reliable multi-view integration remains challenging (Liu et al., 2025; Zhang et al., 2025; Yu et al., 2025). These findings motivate methods that explicitly ground viewpoints and align observations across views, such as pose-aware approaches that use camera pose as a geometric anchor for multi-view fusion and perspective-consistent reasoning (Chen et al., 2024a; Liao et al., 2025).

2.2 PERSPECTIVE-TAKING AND ALLOCENTRIC REASONING IN MLLMS

Beyond reasoning within a single image, many embodied and multi-view tasks require perspective taking, where the model answers questions from an alternative viewpoint that is unobserved and specified in natural language. This setting calls for allocentric scene understanding and reliable viewpoint grounding, yet current MLLMs can be brittle under perspective shifts even with multiple context images (Ma et al., 2023; Yin et al., 2025; Yang et al., 2025b; Yeh et al., 2025; Li et al., 2025). Recent approaches attempt to bridge this gap via mental imagery or generative rollouts that synthesize missing observations at inference time (Lee et al., 2025; Yang et al., 2025c; Cao et al., 2025). While effective in some cases, these pipelines typically do not explicitly ground the language-specified viewpoint to a concrete target pose, and instead rely on searching over candidate motions and viewpoints. As a result, synthesized observations may drift from the intended viewpoint described in language, producing evidence that is misaligned with the target perspective. Moreover, searching over many candidates can be computationally expensive, making such approaches less suitable for applications that require timely, interactive feedback.

2.3 LANGUAGE-GROUNDED VIEWPOINT IMAGINATION

A key challenge underlying perspective taking is to imagine a faithful observation from a language-specified, unobserved viewpoint and use it to support cross-view alignment and spatial reasoning. Existing approaches largely fall into two lines. One line relies on pose-free image generation and editing models to synthesize a new view directly from text and contextual images (Team et al., 2023; Wu et al., 2025; Achiam et al., 2023). However, such generations provide no explicit control over camera information and may not be reliably 3D-consistent under perspective shifts, making the imagined evidence brittle for viewpoint-sensitive reasoning. The other line uses pose-conditioned novel-view synthesis models, which can produce geometrically consistent renderings when a target pose is given (Zhou et al., 2025; Jin et al., 2024; Yu et al., 2021; Charatan et al., 2024; Chen et al., 2024b; Wu et al., 2024; Gao et al., 2024), but they do not address the crucial missing step in our setting: mapping a natural-language viewpoint description to the target camera pose. CAMCUE bridges these two lines by learning to predict the target camera pose from language and using it as an explicit geometric anchor for token-level fusion and image imagination.

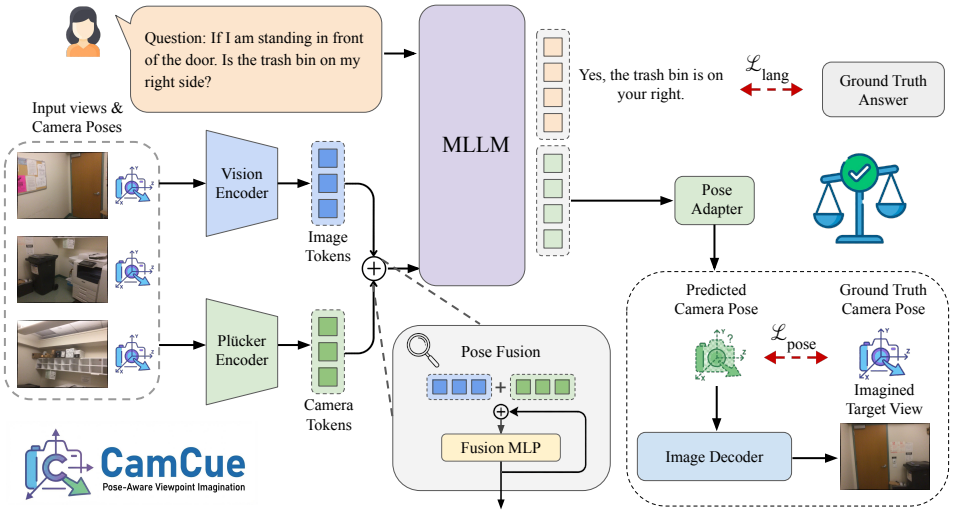


Figure 2: CamCue encodes multi-view images with their camera poses, fuses pose-aware visual tokens, and passed to an MLLM to jointly answer the question and predict the target camera pose. The predicted pose can condition an image decoder to synthesize an imagined target view, which is fed back as additional evidence.

3 METHOD

In this section, we introduce CAMCUE, a pose-aware multi-image framework for perspective-shift spatial reasoning. Figure 2 provides an overview of the CAMCUE pipeline. Given a text prompt T that contains a natural-language description of a target perspective and a question, together with a set of V contextual images $\mathcal{I} = \{I_i\}_{i=1}^{V}$ and their associated camera poses $\mathcal{P} = \{P_i\}_{i=1}^{V}$, CAMCUE predicts the answer under the specified target perspective. We first present the CAMCUE model architecture in Sec. 3.1. We then describe the construction of CAMCUE-DATA in Sec. 3.2.

3.1 ARCHITECTURE

Plücker encoder As shown in Figure 2, each contextual view provides camera extrinsics $\mathbf{C}_i \in \mathbb{R}^{4\times 4}$ and intrinsics $\mathbf{K}_i$.

Following prior work (Jiang et al., 2025), we transform $(\mathbf{C}_i, \mathbf{K}_i)$ into a pixel-aligned Plücker ray map

$$\mathbf{R}_i = Plücker(\mathbf{C}_i, \mathbf{K}_i) \in \mathbb{R}^{H\times W\times 6}, \tag{1}$$

which represents the camera pose information as dense rays aligned with image pixels.

We then encode $\mathbf{R}_i$ into patch-aligned camera tokens

$$Z_i = E_{\text{pose}}(\mathbf{R}_i) \in \mathbb{R}^{S\times d}, \tag{2}$$

where $S = H_p W_p$ is the number of patch tokens under the backbone's canonical resolution, with $H_p = H'/p$ and $W_p = W'/p$ for patch size p.

E_{pose} is a lightweight Plücker encoder that follows the same patchification and spatial aggregation scheme as the vision backbone. Specifically, E_{pose} follows the same tokenization pipeline as the vision encoder: it first resizes $\mathbf{R}_i$ to the backbone's canonical resolution, then patchifies it and applies a patch embedding to convert each local ray patch into a d-dimensional token. This yields a patch-aligned token grid that is spatially aligned with the image patch tokens for subsequent fusion.

Pose-aware token fusion. Given the image patch tokens $X_i \in \mathbb{R}^{S\times d}$ from the vision backbone and the corresponding Plücker camera tokens $Z_i \in \mathbb{R}^{S\times d}$, we fuse pose information into the visual representation in a patch-aligned manner. We concatenate tokens at the same patch index and apply a lightweight MLP projection to produce a residual update:

$$\tilde{X}_i = X_i + W\,[Z_i; X_i], \tag{3}$$

Table 1: Comparison between CAMCUE and related datasets. *Camera Pose* is marked when per-view pose is included in the released benchmark. *Pose Desc.* is marked when the benchmark includes perspective taking data, and *Target View* denotes the corresponding ground-truth image.

Dataset	Camera Pose	Multi-img	Pose Desc.	Target View
ScanQA (Azuma et al., 2022)	✓	✗	✗	✗
SQA3D (Ma et al., 2023)	✓	✗	✓	✗
MMSI (Yang et al., 2025b)	✗	✓	✓	✗
MindCube (Yin et al., 2025)	✗	✓	✓	✗
VSI-Bench (Yang et al., 2025a)	✗	✗	✓	✗
All-Angles (Yeh et al., 2025)	✗	✓	✓	✗
SAT (Ray et al., 2024)	✗	✓*(1/2)*	✓	✗
CAMCUE (Ours)	✓	✓	✓	✓

where $[\cdot;\cdot]$ denotes feature-wise concatenation and $W \in \mathbb{R}^{d\times 2d}$. This design preserves the backbone token layout while injecting per-patch geometric cues, yielding fused tokens $\tilde{X}_i \in \mathbb{R}^{S\times d}$ for subsequent multi-view reasoning.

Target pose prediction. As shown in Figure 2 (the *Pose Adapter* branch), given the fused multi-view scene tokens $\tilde{X} \in \mathbb{R}^{T_{\text{vis}}\times d}$ and the text hidden states $H \in \mathbb{R}^{T_{\text{text}}\times d}$, we predict the target camera pose with a query-based cross-attention head. Concretely, we introduce N learnable query vectors $Q_0 \in \mathbb{R}^{N\times d}$ that attend to the concatenated sequence of text and visual tokens:

$$Y = \mathrm{Attn}\Big(Q_0,\ [H;\tilde{X}],\ [H;\tilde{X}]\Big) \in \mathbb{R}^{N\times d}, \tag{4}$$

where $\mathrm{Attn}(\cdot)$ is multi-head attention and $[\cdot;\cdot]$ denotes concatenation along the token dimension. We then project each attended query with a linear projection to obtain pose query tokens

$$U = \psi(Y) \in \mathbb{R}^{N\times d_q}. \tag{5}$$

We set $N = 16$ and map the 16 pose query tokens to a camera-to-world matrix:

$$\hat{\mathbf{C}}_{\text{tgt}} = \mathrm{reshape}(g(U)) \in \mathbb{R}^{4\times 4}, \tag{6}$$

where $g(\cdot)$ produces one scalar per token.

Answer generation. Our model produces the language answer and the target pose prediction in a single pass. Concretely, the model autoregressively generates a response sequence that begins with a pose slot segment and then outputs the final text answer. At inference time, we optionally use the predicted camera pose to synthesize an imagined target observation with an image decoder; in our implementation, we use LVSM (Jin et al., 2024). The synthesized image is then treated as additional visual evidence and provided to the MLLM to answer the question again, yielding an evidence-enhanced prediction.

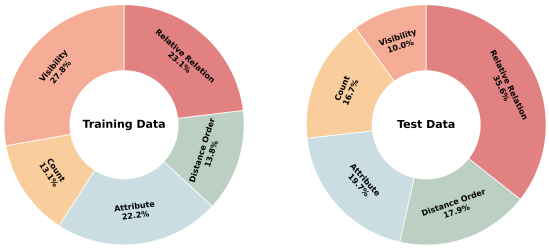
Training objective. We train the model with a weighted sum of language modeling and pose regression losses:

$$\mathcal{L} = \lambda_{\text{lang}}\mathcal{L}_{\text{lang}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}}. \tag{7}$$

$\mathcal{L}_{\text{lang}}$ is the standard cross-entropy loss on text output. $\mathcal{L}_{\text{pose}}$ supervises the predicted target-view camera extrinsics $\hat{\mathbf{C}}_{\text{tgt}}$ with the ground-truth extrinsics $\mathbf{C}_{\text{tgt}}$:

$$\mathcal{L}_{\text{pose}} = \mathrm{MSE}(\hat{\mathbf{t}}, \mathbf{t}) + \mathrm{MSE}(\hat{\mathbf{R}}, \mathbf{R}), \tag{8}$$

where $(\mathbf{R}, \mathbf{t})$ and $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ denote the rotation and translation components extracted from $\mathbf{C}_{\text{tgt}}$ and $\hat{\mathbf{C}}_{\text{tgt}}$, respectively, i.e., $\mathbf{R} = (\mathbf{C}_{\text{tgt}})_{1:3,1:3}$, $\mathbf{t} = (\mathbf{C}_{\text{tgt}})_{1:3,4}$ and similarly $\hat{\mathbf{R}} = (\hat{\mathbf{C}}_{\text{tgt}})_{1:3,1:3}$, $\hat{\mathbf{t}} = (\hat{\mathbf{C}}_{\text{tgt}})_{1:3,4}$.



(a) Training data (b) Test data

Figure 3: QA type distribution in training and test splits.

3.2 Data Construction

We construct CamCue-Data to evaluate and train perspective-shift spatial reasoning under a realistic multi-view setting, where models must answer questions from a new camera pose position described in language while grounding on a sparse set of contextual observations. As summarized in Table 1, existing resources typically cover only a subset of these requirements. Many prior situated question answering datasets assume access to a complete 3D scene representation, rather than sparse multi-view observations. Meanwhile, recent multi-image benchmarks may include viewpoint language as cues embedded in questions, but often do not provide an explicit, detailed target-view pose description paired with camera pose information.

To fill this gap, we curate CamCue-Data with 27,668 training QA pairs and 508 test QA pairs. Each example pairs sparse multi-view observations and their camera poses with a diverse, detailed natural-language description of a novel target viewpoint. The test set further includes expert-annotated viewpoint descriptions, allowing us to assess robustness and performance in realistic interactive scenarios. Example data samples are provided in Appendix A.2.

Data Collection & Preprocessing. We derive training and test QA pairs from the ScanNet training and test splits, respectively. For each scene, we form a multi-view sample by choosing one target view and selecting four contextual views. We first filter candidate contextual views by a moderate translation range to the target and require sufficient viewpoint change so that contextual views are not near-duplicates. We then select four contextual views to ensure the target view is well supported by the contextual observations, using depth-based visibility checks to verify that the contextual views jointly cover what would be seen from the target viewpoint. We keep only target-context groups that pass this visibility criterion and discard redundant groups that are too similar in pose. Qualitative examples are shown in Figure 5, and details are deferred to the Appendix A.1.

Target Pose Descriptions. Each data sample includes a target pose description that specifies the novel camera location and orientation. We generate these descriptions with GPT-4.1 (Achiam et al., 2023) and diversify the phrasing to improve robustness. We use three description styles: layout-anchored descriptions place the camera with respect to the overall room layout; landmark-relative descriptions specify the viewpoint via relative relations among objects; and object-centric descriptions center the viewpoint around a dominant furniture landmark. To evaluate whether models generalize to human-written pose descriptions, we additionally collect expert-written descriptions for the test split. Annotators are instructed to describe the camera position and viewing direction of the target image in clear natural English and to avoid ambiguous phrasing or references so that the described viewpoint is uniquely identifiable.

Question Construction. Given the contextual views, target pose description, and target view, we curate QA pairs that require answering from the described target perspective. Following prior spatial reasoning benchmarks, we organize questions into five types: *Attribute*, *Count*, *Distance Order*, *Relative Relation*, and *Visibility*. Attribute queries an object's attribute; Count queries the number of instances; Distance Order compares which object is closer to the camera; Relative Relation queries relative spatial relations between objects; Visibility queries whether an object is visible from the target viewpoint; see Appendix A.2 for concrete examples. The distribution across training and test splits is shown in Figure 3. We generate QA pairs with GPT-4.1, and manually review the test split to ensure questions are unambiguous and answers are correct.

Table 2: Main results on perspective-shift reasoning.

Model	Overall (Avg.)	Attribute	Visibility	Distance Order	Relative Relation	Count
Qwen2.5-VL-7B	71.06	93.00	84.31	71.43	59.73	55.29
+ MindJourney	72.83	92.00	84.31	80.22	65.75	50.59
+ **CamCue**	**80.12**	92.00	88.24	83.52	78.52	60.00
Qwen2.5-VL-3B	67.52	97.00	80.39	62.64	57.05	50.59
+ MindJourney	70.28	95.00	76.47	69.23	64.64	50.59
+ **CamCue**	**75.92**	94.12	82.35	80.43	67.97	63.53
InternVL-2.5-8B	68.11	89.00	76.47	80.22	58.56	52.94
+ MindJourney	74.21	93.00	82.35	85.71	64.64	55.29
+ **CamCue**	**77.36**	94.00	80.39	89.01	72.38	54.12

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset and Benchmarks. We evaluate CAMCUE on perspective-shift reasoning using CAMCUE-DATA, and general multi-image spatial reasoning benchmarks, including MMSI (Yang et al., 2025b) and MindCube (Yin et al., 2025), which do not provide explicit camera pose inputs. This allows us to verify that CAMCUE yields substantial gains on perspective-taking, while preserving general multi-image reasoning performance.

Backbones and baselines. We deploy CAMCUE on top of QwenVL2.5-3B, QwenVL2.5-7B (Bai et al., 2025), and InternVL2.5-8B (Chen et al., 2024c). We compare CAMCUE with the backbone-only setting and also MindJourney (Yang et al., 2025c), a competitive test-time scaling method that calls an external world model to generate auxiliary observations and feeds them back to the MLLM for answering. Although MindJourney is primarily formulated for a single-reference-view setting, it uses Stable Virtual Camera (SVC) (Zhou et al., 2025) as the underlying world model, which can also condition on multiple contextual images with known camera poses to synthesize an imagined observation set at inference time. We therefore adapt MindJourney to our multi-view setting by providing the full set of contextual views and poses to SVC.

Training setup. We use mixed-data training by interleaving MindCube training examples without camera poses, mixing one MindCube example for every five CAMCUE training data. For such samples we omit pose loss and optimize only $\mathcal{L}_{\text{lang}}$, enabling the model to fall back to standard multi-image inference while still leveraging pose when provided. We fine-tune all backbones using LoRA with a cosine learning-rate scheduler and a batch size of 8. We train the pose-related modules with a learning rate of 5×10^{-5} and the LoRA parameters with a learning rate of 1×10^{-5}, using a warmup ratio of 0.03. More details are in the Appendix B.

4.2 EXPERIMENT RESULTS

Main results on perspective-shift reasoning. Table 2 reports accuracy on CAMCUE-DATA. CAMCUE yields consistent gains across all three backbones, with the largest improvements on viewpoint-sensitive categories such as visibility, distance order, and relative relation. Notably, MindJourney improves over backbone-only inference, confirming the benefit of inference-time imagination for perspective reasoning. However, it remains substantially below CAMCUE. A key reason is that MindJourney typically explores a set of navigational rollouts (e.g., turn left/right, move forward) to collect additional synthesized observations, rather than grounding the natural-language viewpoint description to a single explicit target viewpoint. As a result, the explored views may be informative but are not guaranteed to coincide with the queried perspective, motivating our explicit viewpoint grounding and pose-conditioned target-view synthesis.

Results on General Multi-image Benchmarks We further evaluate general multi-image spatial reasoning on MMSI and the MindCube Tiny benchmark. Since these benchmarks do not provide camera pose information, we evaluate CAMCUE without target-view synthesis, where the model answers directly from the given contextual images. As shown in Table 3, CAMCUE improves overall accuracy on both benchmarks, indicating that our training does not compromise general multi-image reasoning and can transfer beyond pose-supervised settings.

Table 3: **Results on MindCube Tiny and MMSI with the Qwen2.5-VL-7B backbone.** Pose-Only denotes CamCue pose-only inference without imagined image feedback.

Model	MindCube Tiny Accuracy (%)			
	Overall	Rotation	Among	Around
Base	29.3	38.7	29.5	21.4
Pose-Only	**47.43**	75.00	31.67	63.20

Model	MMSI Accuracy (%)				
	Overall	Position	Attribute	Motion	MSR
Base	25.9	25.9	21.5	30.0	25.8
Pose-Only	**28.8**	29.9	29.2	27.3	26.8

Table 4: **Camera pose estimation accuracy under different viewpoint description sources.** Values are the percentage of samples with rotation/translation error within each threshold.

Viewpoint Description	Rotation Acc. ↑ (%)			Translation Acc. ↑ (%)		
	R@5°	R@10°	R@20°	t@0.1	t@0.3	t@0.5
Synthetic	19.3	35.4	91.5	12.0	62.4	92.9
Human Description	30.1	56.9	100.0	19.5	74.8	95.1

Camera Pose Prediction Table 4 reports target pose prediction accuracy when the desired viewpoint is specified by a natural-language description. We consider two description sources: GPT-4.1 generated descriptions and human expert annotations. Following standard pose-evaluation practice, we report the fraction of samples whose rotation and translation errors fall within each threshold. Overall, CAMCUE achieves high target-pose prediction accuracy from natural language description. Under synthetic descriptions, 91.5% of examples have a camera rotation error below 20°, and 92.9% have a translation error below 0.5; with human-written descriptions, these fractions increase to 100.0% and 95.1%, respectively. We attribute this gap to the fact that expert annotations are typically more detailed and less ambiguous than LLM-generated descriptions, providing clearer cues about the camera position. To qualitatively verify geometric fidelity, Figure 5 visualizes the predicted pose by rendering the scene from the estimated camera and comparing it against the ground-truth target view. The close alignment in both viewpoint and visible content indicates that CAMCUE reliably grounds language to precise camera geometry, providing a dependable basis for downstream perspective-shift reasoning.

4.3 ABLATIONS AND ANALYSIS

Ablation studies. Table 5 analyzes Qwen2.5-VL-7B to disentangle the effects of pose supervision and imagined-view feedback. QA-FT (fine-tuning with QA supervision only) yields only marginal changes from the base model, indicating that answer-only supervision does not reliably teach perspective shifting. In contrast, Pose only model (training with pose supervision but no target-view synthesis at inference) already improves over QA-FT, showing that learning to predict and use camera pose provides a meaningful geometric prior even without imagined images. Building on this, CAMCUE further introduces target-view synthesis and image feedback, leading to a substantial jump across all categories, which confirms the importance of converting viewpoint grounding into concrete visual evidence for answering. Finally, CAMCUE (GT) replaces the imagined view with the ground-truth target view and serves as an oracle upper bound, suggesting additional headroom from improved novel-view synthesis quality.

Faithful Viewpoint Imagination Camera pose is an effective intermediate variable for connecting multi-view contextual understanding with target-view inference. Predicting the target camera pose encourages the model to represent the viewpoint description in an explicit geometric form, which helps relate observations across contextual views under viewpoint change. This geometric anchor can then be used to guide a 3D-aware imagination step, where the synthesized target view is constrained by the predicted camera pose and is therefore better aligned with the ground-truth physical

Table 5: **Ablation study on Qwen2.5-VL-7B backbone.** (1) fine-tuning with QA supervision only (no pose). (2) pose-only inference without imagined image feedback. (3) full pipeline with imagined image feedback. (4) oracle upper bound by replacing the imagined image with the GT target view.

	Method	Overall (Avg.)	Attribute	Visibility	Distance Order	Relative Relation	Count
(0)	Base	71.06	93.00	84.31	71.43	59.73	55.29
(1)	QA-FT	71.26	92.00	78.43	67.03	65.75	58.82
(2)	Pose-Only	72.44	90.00	80.39	80.22	65.19	54.12
(3)	CamCue	80.12	92.00	88.24	83.52	78.52	60.00
(4)	CamCue (GT)	**87.20**	98.00	98.04	90.11	83.43	72.94

Table 6: **Comparison with a pose-free image-generation baseline.** All methods use the same 7B MLLM backbone for answering (Qwen2.5-VL-7B). **Nano Banana / Nano Banana Pro** synthesize the target view from multi-view context images and a viewpoint description, and the generated image is fed back to the backbone for QA.

Method	Avg.	Attr.	Vis.	Dist.	Rel.	Cnt.
Base	71.06	93.00	84.31	71.43	59.73	55.29
Nano Banana	66.73	82.00	76.47	75.82	58.56	50.59
Nano Banana Pro	76.77	95.00	94.12	82.42	69.61	54.12
CamCue	**80.12**	92.00	88.24	83.52	78.52	60.00

scene than pose-free generation that must infer geometry and viewpoint implicitly from images and language.

This distinction is critical when the imagined observation is used as evidence for answering. The generated target view must remain faithful and stick to the underlying physical environment, otherwise spurious details can mislead downstream reasoning. Despite being a strong generator, Nano Banana (Team et al., 2023) frequently hallucinates or drifts in layout, viewpoint, and object configurations, and can even perform worse than directly answering from the contextual views. As shown in Table 6, Nano Banana underperforms the Base setting by 4.33%. A stronger variant, Nano Banana Pro, produces more faithful images on average and yields some improvement, but its pose-free generations are still not reliably grounded and remain unstable in challenging cases. In contrast, CAMCUE uses the predicted camera pose as an explicit geometric anchor to constrain imagination, leading to a substantial gain and more dependable evidence for perspective-shift reasoning.

These failures are visible in Figure 5: although Nano Banana often preserves the coarse semantics of the scene, its pose-free generations can drift from the ground-truth environment, such as altering global layout and background structures (Example 1–2), or keeping the scene content but misestimating the viewing direction and even changing object configuration (Example 3–4).

Efficiency Table 7 compares inference-time cost per example. CAMCUE remains efficient in practice, it predicts the target pose in a single forward pass, and synthesizes the imagined observation with a feed-forward image decoder, enabling fast end-to-end feedback for interactive use. In contrast, MindJourney shows substantially higher latency because it performs test-time scaling via iterative search over multiple candidate rollouts and repeatedly queries the world model and VLM to aggregate evidence.

5 CONCLUSION

We propose CAMCUE, a pose-aware framework that equips MLLMs with explicit viewpoint grounding for multi-image spatial reasoning. CAMCUE predicts the target camera pose from multi-view observations and a natural language viewpoint description, and uses this pose as an anchor to support target-view inference. Building on the predicted camera pose, our pipeline can optionally synthesize a target-view observation as additional evidence, providing a more faithful form of viewpoint imagination for downstream reasoning tasks. Across experiments, CAMCUE consistently improves overall performance over strong baselines, while remaining efficient in practice, substantially reducing inference-time cost compared to prior methods that rely on iterative test-time search.

Table 7: **Inference-time cost per example.**

Method	CamCue	Nano Banana	MindJourney
Time (s)	1.453	35.1	256.6

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19129–19139, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. *arXiv preprint arXiv:2512.07733*, 2025.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Junyi Chen, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Where am i and what will i see: An auto-regressive model for spatial localization and view prediction. In *The Thirteenth International Conference on Learning Representations*, 2024a.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025.

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

Kang Liao, Size Wu, Zhonghua Wu, Linyi Jin, Chao Wang, Yikai Wang, Fei Wang, Wei Li, and Chen Change Loy. Thinking with camera: A unified multimodal model for camera-centric understanding and generation. *arXiv preprint arXiv:2510.08673*, 2025.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *ArXiv*, abs/2511.15722, 2025. URL https://api.semanticscholar.org/CorpusID:283109806.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes, 2023. URL https://arxiv.org/abs/2210.07474.

Tobias Meilinger, Alain Berthoz, and Jan M. Wiener. The integration of spatial information across different viewpoints. *Memory & Cognition*, 39(6):1042–1054, 2011. ISSN 1532-5946. doi: 10.3758/s13421-011-0088-x. URL https://doi.org/10.3758/s13421-011-0088-x.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Fengyun Wang, Sicheng Yu, Jiawei Wu, Jinhui Tang, Hanwang Zhang, and Qianru Sun. 3d question answering via only 2d vision-language models. *arXiv preprint arXiv:2505.22143*, 2025.

Ranxiao Wang. Theories of spatial representations and reference frames: What can configuration errors tell us? *Psychonomic bulletin review*, 19:575–87, 05 2012. doi: 10.3758/s13423-012-0258-2.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21551–21561, 2024.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

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

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025b.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025c.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Rouyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *ArXiv*, abs/2504.15280, 2025. URL https://api.semanticscholar.org/CorpusID:277955366.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *ArXiv*, abs/2509.18905, 2025. URL https://api.semanticscholar.org/CorpusID:281496332.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture, 2025. URL https://arxiv.org/abs/2509.02359.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

APPENDIX

A DATA CURATION

A.1 VIEW SELECTION DETAILS

Algorithm 1 Multi-view group selection with depth-based visibility.

$\mathcal{G} \leftarrow \emptyset$
for each target view t **do**
  $\mathcal{C} \leftarrow \{c \neq t \mid \mathrm{PoseFilter}(t, c)\}$
  if $|\mathcal{C}| < 4$ **then**
    continue
  end if
  $P \leftarrow \mathrm{TargetSamples}(t)$
  $\mathcal{S} \leftarrow \emptyset;\ M \leftarrow \emptyset$
  for $r = 1$ to 4 **do**
    Choose $c^\star \in \mathcal{C} \setminus \mathcal{S}$ maximizing $|\mathrm{Vis}(P, t, c) \setminus M|$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{c^\star\};\ M \leftarrow M \cup \mathrm{Vis}(P, t, c^\star)$
  end for
  if $|M|/|P| < \gamma$ **then**
    continue
  end if
  if $\mathrm{Redundant}(t, \mathcal{S}, \mathcal{G})$ **then**
    continue
  end if
  $\mathcal{G} \leftarrow \mathcal{G} \cup \{(t, \mathcal{S})\}$
end for
return $\mathcal{G}$

Algorithm 2 Depth-based visibility test used in $\mathrm{Vis}(P, t, c)$.

Procedure $\mathrm{Vis}(P, t, c)$
$V \leftarrow \emptyset$
for each sample $p \in P$ **do**
  Back-project p in target view using (D_t, K, C_t) to obtain a 3D point X
  Project X into context view c to get $(u, v, \hat{z})$
  if (u, v) is in bounds **and** $\hat{z} \leq D_c(u, v) + \epsilon$ **then**
    $V \leftarrow V \cup \{p\}$
  end if
end for
return V

Parameters. For candidate context filtering, we keep frames within a moderate translation range to the target, $d \in (0.4, 2.5)$ meters, and avoid near-duplicate views using a distinctness heuristic: $d > 0.6$ m, or $\Delta\theta > 15°$. We then greedily select 4 context views to maximize depth-based visibility coverage (Algorithm 2), and keep a target–context group only if the overall coverage satisfies $|M|/|P| \geq \gamma$ with $\gamma = 0.80$. Finally, we remove redundant groups by discarding a candidate target view if it is too similar to an existing one, using a translation threshold $\tau_t = 0.5$ m and a rotation threshold $\tau_\theta = 45°$.

A.2 DATA SAMPLES

Figure 4 shows three representative examples from CAMCUE-DATA. Each datapoint consists of four contextual images, a natural-language description specifying a target viewpoint, the target-view image, and QA pairs that must be answered from the described target perspective.

Figure 4: Data Samples from CamCue

Example 1 (Kitchen scene). **Target-view description:** *The camera is to the right of the stove and counter, near the corner where the counter meets the wall. It is facing the wall with windows, with the stove and counter on the left side of the view and the heater on the right side of the view.*

Count: *How many windows are visible in this image?*

Options: (A) 1 (B) 2 (C) 3 (D) 4. **Answer:** (B).

Visibility: *Can you see the fire extinguisher from this viewpoint?*

Options: (A) Yes (B) No. **Answer:** (A).

Example 2 (Auditorium). **Target-view description:** *The camera is to the right of the rows of auditorium chairs, close to the wall with the handrail. It is aimed toward the wall with the handrail and door, with the chairs on the left side of the view and the wall on the right side of the view.*

Relative relation: *Where is the handrail located relative to the seats in this image?*

Options: (A) Behind (B) In front of (C) Left of (D) Right of. **Answer:** (A).

Attribute: *What is the most likely material of the handrail visible on the wall?*

Options: (A) Wood (B) Metal (C) Plastic (D) Glass. **Answer:** (A).

Example 3 (Gym). **Target-view description:** *The camera is placed to the right of the treadmills, near where the mat meets the wooden floor. It is facing toward the dumbbell rack on the left side of the view and the curved wall on the right side of the view.*

Distance order: *Which is closer to you, the treadmill or the dumbbell rack?*

Options: (A) The treadmill (B) The dumbbell rack. **Answer:** (A).

B Training Hyperparameters

Tables 8–9 summarize the key training hyperparameters used in our experiments. We fine-tune the language backbone with LoRA while jointly training the pose-related modules. The overall objective is a weighted sum of language modeling loss and pose regression loss with $\lambda_{\text{lang}} = 1.0$ and $\lambda_{\text{pose}} = 0.2$.

Table 8: Training hyperparameters.

Hyperparameter	**Qwen2.5-VL (3B/7B)**	**InternVL2.5 (8B)**
#GPUs	4	4
Per-device batch size	2	2
Effective batch size	8	8
Training steps	9000	9000
Scheduler	Cosine	Cosine
Warmup ratio	0.03	0.03
Geometry LR	5×10^{-5}	5×10^{-5}
Language (LoRA) LR	1×10^{-5}	1×10^{-5}

Table 9: Model adaptation and loss weights.

Setting	**Qwen2.5-VL (3B/7B)**	**InternVL2.5 (8B)**
LoRA rank r	16	16
LoRA α	64	64
LoRA dropout	0.10	0.10
LoRA target modules	{q,k,v,o proj, gate/up/down proj}	
Pose loss weight λ_{pose}	0.2	0.2
Language loss weight λ_{lang}	1.0	1.0

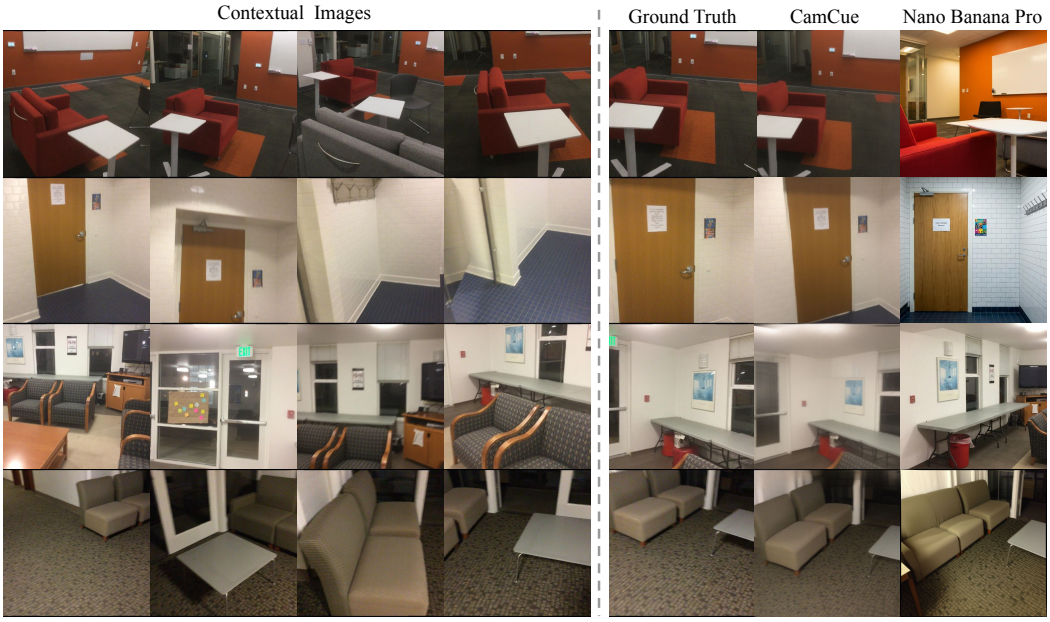


Figure 5: Qualitative comparison of imagined target views.

C Render Examples

Figure.5 shows additional examples of imagined target views produced from the same contextual images and target-viewpoint description. We compare the ground-truth target view with CamCue and Nano Banana Pro.

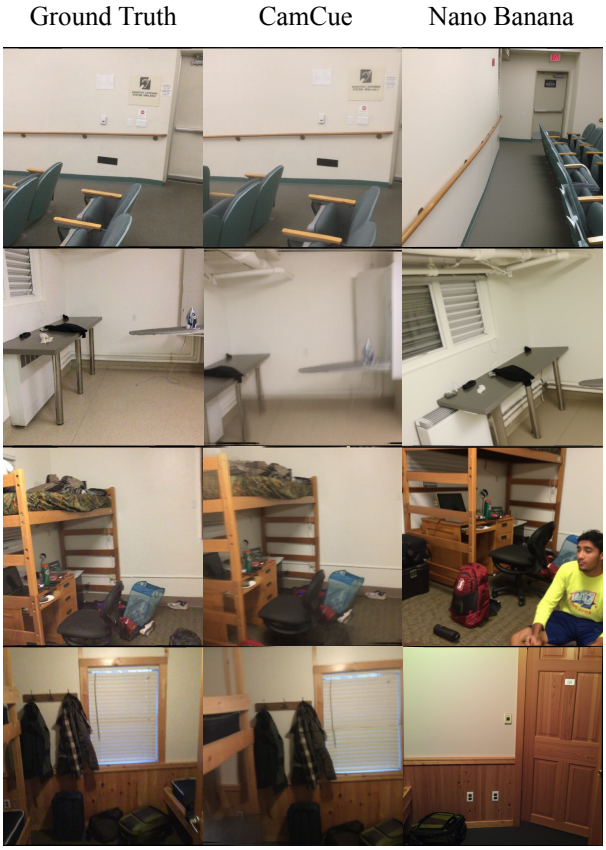


Figure 6: Additional qualitative examples and failure cases.

D ADDITIONAL QUALITATIVE EXAMPLES AND FAILURE CASES

In Figure 6, we compare the ground-truth target view with synthesized views from CAMCUE and Nano Banana. CAMCUE generally preserves the scene layout but may produce blurred renderings in some cases. In contrast, Nano Banana often generates visually sharp images but may exhibit inaccurate viewpoint estimation, and can introduce changes to the environment (e.g., modifying object placements or adding/removing scene elements), which breaks geometric and physical consistency for spatial reasoning.