ViewFusion: Structured Spatial Thinking Chains for Multi-View Reasoning

Xingjian Tao[1], Yiwei Wang[3], Yujun Cai[4], Yifan Song[1], and Jing Tang[1,2⋆]

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] University of California, Merced, Merced, CA, USA
[4] The University of Queensland, Brisbane, QLD, Australia
taoxj2001@outlook.com, jingtang@ust.hk

Abstract. Multi-view spatial reasoning remains difficult for current vision-language models. Even when multiple viewpoints are available, models often underutilize cross-view relations and instead rely on single-image shortcuts, leading to fragile performance on viewpoint transformation and occlusion-sensitive cases. We present ViewFusion, a two-stage framework that explicitly separates cross-view spatial pre-alignment from question answering. In the first stage, the model performs deliberate spatial pre-thinking to infer viewpoint relations and spatial transformations across views, forming an intermediate workspace that goes beyond a simple re-description. In the second stage, the model conducts question-driven reasoning conditioned on this workspace to produce the final prediction. We train ViewFusion with synthetic reasoning supervision followed by reinforcement learning using GRPO, which improves answer correctness while stabilizing the intended two-stage generation behavior. On MMSI-Bench, ViewFusion improves accuracy by 5.3% over Qwen3-VL-4B-Instruct, with the largest gains on examples that require genuine cross-view alignment. Our code is available at https://github.com/taoxj2001/ViewFusion.

Keywords: Vision Language Models · Spatial Reasoning · Reinforcement Learning

1 Introduction

Recent advances in vision-language models have enabled machines to understand and reason about visual content alongside natural language. Models such as LLaVA [19, 22, 23], Flamingo [1], Gemini [30], and Qwen-VL [2, 4, 33] have expanded the scope of multimodal intelligence from conventional tasks such as visual question answering, image captioning, and document understanding to broader real-world applications, including chart and video understanding, visual grounding, embodied decision-making, and GUI agents that interact with web pages [7, 14, 24, 26, 28, 29]. These developments indicate a transition from passive

⋆ Corresponding author: Jing Tang.

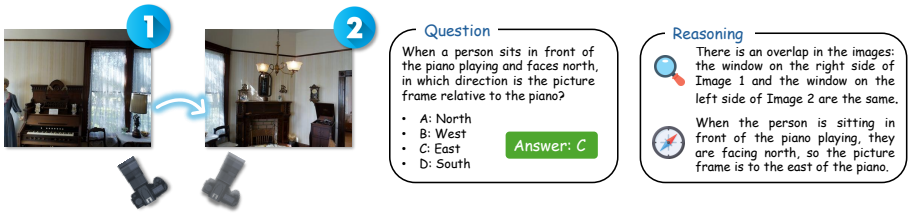

Fig. 1: An example of multi-view spatial reasoning. Given two images captured from different viewpoints, the model must align shared visual cues across views to infer the viewpoint relationship and answer a direction-based question (e.g., locating the picture frame relative to the piano)

visual understanding toward action-oriented multimodal reasoning. However, multi-view spatial reasoning, the ability to align spatial information across different viewpoints and reason about 3D scene structure, remains a fundamental challenge that current models struggle to solve reliably.

The core difficulty lies in cross-view spatial alignment. When presented with multiple images of the same scene from different viewpoints, models must not only recognize objects and their attributes within each view, but also establish spatial correspondences across views: How has the camera moved? Which objects correspond under viewpoint change? How do occlusions evolve as perspective shifts? These cross-view relations are essential for answering questions that require genuine multi-view understanding, yet they remain largely implicit in current reasoning approaches.

Consider a concrete example illustrated in Figure 1. Given two living-room images captured from different viewpoints, a question asks: "When a person sits in front of the piano playing and faces north, in which direction is the picture frame relative to the piano?" In the first view, the piano is seen beside a tall window and the picture frame is not clearly localized; in the second view, the scene is observed from a shifted angle where the window and wall decor provide an overlap across views and the picture frame becomes visible on the corresponding wall. To answer correctly, the model must align these shared cues to infer the relative viewpoint change and then map the picture frame's position into the question's north-referenced coordinate frame. Without proper cross-view alignment, the model may mis-associate landmarks across images or reason from a single view, resulting in an incorrect direction even when each individual image is well described.

In practice, many existing approaches [6,23,31] still rely heavily on single-view cues and superficial correlations, which leads to brittle behavior and noticeable performance drops when multiple viewpoints must be aligned and complementary evidence integrated to resolve ambiguities. This limitation also persists when adopting reinforcement learning as a post-training strategy. While RL methods such as Group Relative Policy Optimization (GRPO) can improve task-level performance from model rollouts [12] and often encourage more elaborate de-

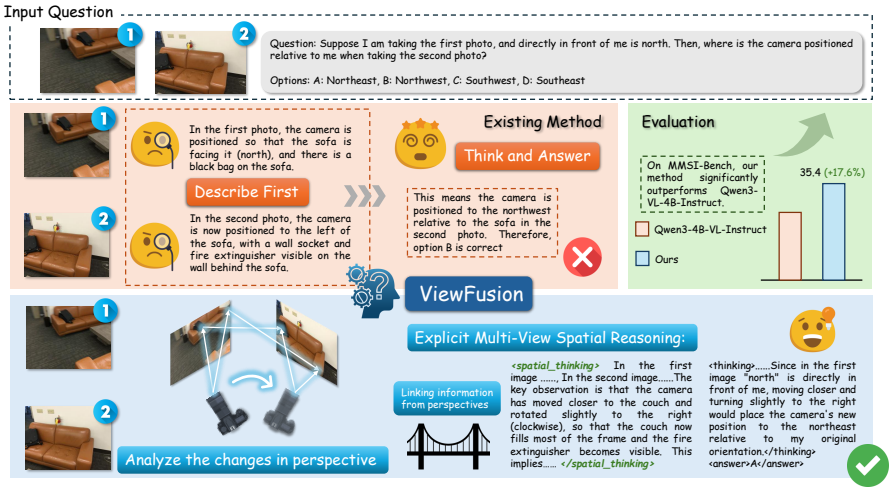

Fig. 2: Overview of ViewFusion for multi-view spatial reasoning. Given a multi-view question (left), existing "describe-first" or direct "think-and-answer" paradigms often produce view-local descriptions and then shortcut to answering without establishing correct cross-view spatial relations, leading to errors (top). ViewFusion instead performs explicit multi-view spatial pre-thinking to link perspectives and infer viewpoint transformations across images before question solving (bottom), yielding more reliable reasoning and correct predictions.

liberation, they do not necessarily induce correct cross-view spatial alignment. In our preliminary study, models trained with vanilla GRPO frequently exhibit shortcut behaviors: they begin solving the question before integrating the full multi-view context, or rely predominantly on a single view while treating other images as incidental. Consequently, the resulting reasoning may appear detailed but remains grounded in an incomplete cross-view spatial model, and providing more views does not reliably translate into better multi-view reasoning. In some cases, additional views can even introduce noise that amplifies shortcut learning and destabilizes intermediate reasoning.

To address the persistent difficulty of establishing correct cross-view spatial relations, we propose ViewFusion as shown in Figure 2, a simple but effective two-stage "think twice" paradigm for multi-view spatial reasoning. ViewFusion explicitly separates spatial pre-thinking from question answering. In the first stage, the model performs deliberate spatial pre-thinking to infer viewpoint relations and spatial transformations across views and organize them into an intermediate workspace. In the second stage, the model conducts question-driven reasoning conditioned on this workspace to produce the final answer. The core principle is to make cross-view alignment a deliberate first step, rather than an implicit byproduct of answering. We train ViewFusion by first performing supervised fine-tuning with synthesized reasoning traces that reflect this two-stage

protocol, and then applying reinforcement learning with GRPO to further align the model toward correct answers while stabilizing consistent two-stage behavior and reliable generation.

Empirically, ViewFusion achieves strong experimental results on MMSI-Bench [44], improving accuracy by 5.3% over Qwen3-VL-4B-Instruct, with particularly large gains on examples that require genuine cross-view alignment. We compare ViewFusion with several popular spatial reasoning baselines and observe consistent improvements across settings. Our ablation studies further validate the reliability of each component in the proposed design. Notably, ViewFusion also outperforms Qwen3-VL-4B-Thinking, suggesting that explicitly enforcing a two-stage pre-thinking protocol yields benefits beyond simply encouraging longer deliberation.

In this paper, our contributions can be summarized as follows:

- We diagnose a key failure mode in current multi-view spatial reasoning with MLLMs, including RL-trained models: they often fail to align cross-view spatial information and instead rely on shortcut behaviors that underuse the available viewpoints.
- We introduce ViewFusion, a two-stage "think twice" paradigm that explicitly separates cross-view spatial pre-thinking from question solving, together with a training recipe that combines synthesized reasoning supervision and GRPO-based reinforcement learning to stabilize this behavior.
- We provide comprehensive experimental evidence on MMSI-Bench, including comparisons with strong and widely used baselines (notably Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking) and ablations that verify the contribution of each component.

2 Related Work

2.1 Reinforcement Learning for MultiModal Large Language Models Reasoning

Reinforcement learning (RL) and preference optimization have become increasingly popular for improving the reasoning quality and behavioral alignment of multimodal large language models (MLLMs) [15, 27, 40, 47, 48]. Beyond outcome-level supervision, recent reasoning-oriented RL methods aim to provide richer training signals that shape the structure of multimodal reasoning. For example, Insight-V [8] leverages a multi-agent setup to select and learn from self-generated reasoning trajectories, while R1-VL [49] introduces step-wise GRPO with dense rule-based rewards to improve multimodal reasoning paths. This GRPO-style training has also been extended to video reasoning, including Video-R1 [10], Video-RTS [35], and Video-STR [34]. In parallel, several works encourage models to "observe first" by introducing explicit context descriptions or observation stages, such as HumanOmniV2 [41], Visionary-R1 [39], and Observe-R1 [13]. However, these observation steps are typically realized as descriptive summaries of the visual input, and do not explicitly induce the model to reason about relationships

across multiple views. Overall, existing RL-based approaches can encourage stronger deliberation and better alignment in MLLMs, yet designing rewards and training protocols that explicitly promote cross-view spatial consistency remains an open challenge for multi-view reasoning.

2.2 Spatial reasoning with MLLMs

Spatial reasoning has emerged as a key frontier for MLLMs, aiming to move beyond object recognition toward understanding relative position, orientation, viewpoint transformation, and occlusion-aware relations in 3D scenes. A growing body of work seeks to strengthen spatial reasoning in MLLMs by improving grounding and spatial representations, for example via spatially aware instruction tuning and curated supervision [5, 6, 9, 11, 17, 21, 31, 32, 37, 46, 51]. More recently, Visual Spatial Tuning [42] trains vision-language models with large-scale spatial perception and reasoning data, producing notably stronger spatial reasoning performance and improved generalization across spatial benchmarks [42].

To systematically evaluate these advances under multi-view inputs, several benchmarks have been proposed [16, 50]. MMSI-Bench [44] focuses on multi-view, multi-image spatial intelligence and includes problems that require aligning evidence across views rather than solving from a single snapshot. MindCube [45] probes whether models can construct and manipulate a coherent "mental model" of the scene from limited observations, emphasizing perspective taking and spatial consistency under incomplete information. ViewSpatial [18] further stresses viewpoint-dependent spatial localization and cross-view reference frames, revealing substantial generalization gaps when camera viewpoints shift. Collectively, these benchmarks provide increasingly fine-grained diagnostics for multi-view spatial understanding and consistently highlight a key bottleneck: strong single-view perception does not automatically translate into reliable cross-view alignment, leaving significant room for methods that explicitly model and enforce spatial consistency across views.

3 ViewFusion

3.1 Limitations of Reasoning Models under Multi-View Inputs

Existing reasoning paradigms for multi-view inputs often fall short because they do not explicitly infer the spatial relationships between images captured from different viewpoints. Instead of establishing cross-view consistency (e.g., how the camera moves, which objects correspond across views, and how occlusions change), the model frequently treats each image as an independent evidence source and proceeds directly to question answering. This "late fusion" behavior implicitly assumes that the relevant information is already visible and interpretable within a single view, so multi-view inputs are used as weak auxiliary context rather than as complementary observations that must be jointly aligned.

Even when an intermediate "observation" or description step is introduced, it is typically view-local and descriptive, summarizing salient entities within each

frame without reasoning about how these entities transform across viewpoints. Such descriptions often omit the most informative cues for multi-view tasks, including which objects disappear due to occlusion versus leaving the scene, how the relative ordering of background landmarks changes under camera motion, and how scale or perspective distortions affect apparent positions.

As a result, key spatial cues that only emerge through cross-view alignment—such as viewpoint-dependent visibility, relative layout changes, and object re-identification under occlusion are easily missed. This leads to reasoning traces that appear coherent but are grounded in an incomplete spatial model, making the final prediction brittle when the task requires genuine multi-view integration. In our error analysis, these failures frequently manifest as plausible but inconsistent narratives (e.g., treating two views as different locations, confusing left/right after a turn, or matching objects to incorrect counterparts) that cannot be corrected by additional textual deliberation alone.

Motivated by these limitations, we argue that effective multi-view reasoning requires an explicit pre-alignment step that prioritizes cross-view spatial consistency before any question-driven inference. Rather than asking the model to "solve while observing", we enforce a "pre-think then answer" protocol in which cross-view alignment is a deliberate first step. Specifically, the model first infers viewpoint relations and spatial transformations across images to form a consistent intermediate workspace, then performs task-specific reasoning conditioned on this workspace. This decomposition targets the shortcut behaviors observed in existing paradigms by (i) forcing the model to reconcile multi-view evidence before committing to an answer, and (ii) making alignment errors more visible and thus easier to constrain during training. As a result, the model becomes more robust on cases where the answer is only recoverable through genuine multi-view integration, such as questions requiring viewpoint transformation, occlusion-aware reasoning, and cross-view correspondence.

3.2 Training Data Preparation

Our training data is sampled from two open-source multi-view datasets, VST-500K [42] and the MindCube-Trainset [45]. We collect raw question–answer pairs together with their corresponding multi-view images, and organize them into two splits to support our two-stage training pipeline: supervised fine-tuning (SFT) and reinforcement learning (RL). All instances follow a unified multi-image QA format, where each example contains 2–8 images paired with a single question and its answer. Our data construction is motivated by a key challenge in multi-view spatial reasoning: when trained directly on multi-view QA, models can easily develop single-view shortcuts (e.g., relying on a salient cue from one image) and skip the cross-view alignment that is essential for correct spatial inference. To mitigate this, we introduce an explicit spatial pre-thinking stage in the SFT supervision, which encourages the model to first reconcile viewpoint relations and cross-view correspondences before answering.

SFT data (18K). We construct an SFT set containing 18K multi-view instances. For each instance, we rewrite the original rationale into a structured reasoning trace using Qwen3-32B-VL-Instruct. Each trace contains three parts: `<spatial_thinking>`, `<thinking>`, and `<answer>`. The `<spatial_thinking>` part is designed to elicit an explicit pre-thinking process that focuses on establishing spatial relations across views (e.g., viewpoint change and cross-view consistency) before proceeding to question-driven reasoning in `<thinking>` and producing the final prediction in `<answer>`. To ensure training stability and format controllability, we apply strict filtering rules and remove samples that violate the required structure (e.g., missing fields, incorrect ordering, unclosed tags, or malformed outputs). This yields a clean SFT corpus with consistent two-stage reasoning behavior.

RL data (16K). We additionally construct an RL set of 16K instances from the same data sources. Unlike the SFT set, the RL split does not include rewritten reasoning traces; it only retains the multi-view input and the supervision necessary for computing outcome-based rewards (e.g., final answer correctness). This separation allows RL to further align the policy toward task success while avoiding overfitting to specific rationales, and enables us to explicitly study how RL interacts with the proposed two-stage reasoning protocol.

3.3 Training Strategy

Preliminary: SFT and GRPO We briefly review supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO), which serve as the two training stages of our framework. Throughout this paper, we view an MLLM as a conditional generative policy $\pi_\theta(y \mid x)$ that produces an output sequence y (e.g., a reasoning trace and the final answer) given a multi-view input x (e.g., a question paired with multiple images). The first stage, SFT, provides a stable initialization that teaches the model the desired generation behavior under teacher forcing. The second stage, RL with GRPO, optimizes the policy directly from rollouts with task-defined rewards, better matching test-time generation and enabling flexible alignment beyond imitation learning.

Supervised Fine-Tuning (SFT). Given a dataset of multimodal instruction-response pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the input (e.g., text with one or multiple images) and y_i the target output, SFT trains a conditional generative policy $\pi_\theta(y \mid x)$ by maximizing the log-likelihood:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log \pi_\theta(y_i \mid x_i). \tag{1}$$

In practice, y_i may include both intermediate reasoning and the final prediction, which encourages the model to internalize a structured inference procedure. SFT is typically data-efficient and stable, but it may inherit biases from the supervision and may not fully optimize for task-level objectives under free-form generation.

Group Relative Policy Optimization (GRPO). Reinforcement learning further optimizes the policy using a reward function $r(x, y)$ defined on model outputs, allowing direct optimization of task success signals. For each input x, GRPO samples a group of K candidate outputs $\{y^{(k)}\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid x)$, and computes a group-wise baseline to reduce variance:

$$A^{(k)} = r(x, y^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} r(x, y^{(j)}). \tag{2}$$

GRPO then updates θ by maximizing a PPO-style clipped objective with a KL regularizer that anchors the policy to a reference model π_{ref}:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{K} \sum_{k=1}^{K} \min\Big(\rho^{(k)} A^{(k)}, \text{clip}\Big(\rho^{(k)}, 1-\epsilon, 1+\epsilon \Big) A^{(k)} \Big) \right] \\ - \beta \, \mathbb{E}_x [D_{\text{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x))], \end{aligned} \tag{3}$$

where $\rho^{(k)} = \frac{\pi_\theta(y^{(k)}|x)}{\pi_{\theta_{\text{old}}}(y^{(k)}|x)}$, ϵ is the clipping threshold, and β controls the KL strength. Compared to outcome-only optimization, GRPO leverages multiple rollouts per instance and uses relative advantages within each group, which provides a low-variance training signal without requiring an explicit value function. The KL term further constrains the update to remain close to π_{ref}, improving stability and preventing excessive drift during RL fine-tuning.

Two-Stage Optimization We train ViewFusion in two stages. We first perform SFT on the curated reasoning corpus to initialize the model with the desired inference protocol, so that it learns to produce structured multi-view reasoning under teacher forcing. This stage serves as a cold start that stabilizes generation and enforces the expected output structure, preventing early training from collapsing into shortcut answers or malformed formats. In our implementation, we use a learning rate of 1×10^{-5} for SFT.

Starting from this initialization, we then apply RL with GRPO to further optimize the policy with sampled rollouts, which better matches test-time generation and directly reinforces behaviors that lead to correct predictions. This stage primarily aims to strengthen multi-view reasoning ability beyond imitation, by optimizing task-level rewards that encourage correct cross-view alignment and robust answering. We use a smaller learning rate of 1×10^{-6} for RL to ensure stable updates, and sample K=8 trajectories per input instance to compute group-relative advantages. This two-stage pipeline combines the stability and controllability of imitation learning with the flexibility of RL to improve robustness under multi-view inputs.

Reward Design for RL A key challenge in RL for reasoning models is that optimizing only for task correctness can easily induce degenerate behaviors. This is particularly pronounced in multi-view settings, where the reward signal is

typically sparse and does not explicitly constrain how the model should use the available views. As a result, correctness-only training may encourage shortcut strategies, such as committing to an answer based on a single salient view, latching onto spurious correlations, or beginning question solving before establishing cross-view consistency. Moreover, focusing solely on the final answer can destabilize generation behavior: the model may omit required intermediate sections, produce malformed outputs that are hard to parse, or collapse to overly terse responses that provide insufficient reasoning content for reliable multi-stage inference. To address these issues, we design a composite reward that explicitly enforces (i) answer correctness, (ii) format compliance with the intended two-stage protocol, and (iii) a reasonable response length that balances sufficient reasoning with verbosity control.

Answer correctness reward. Since all training instances are multiple-choice questions, we extract the predicted option from the `<answer>` field and use a binary reward:

$$r_{\text{ans}}(x, y) = \mathbb{I}[\text{ans}(y) = \text{gt}(x)]\,, \tag{4}$$

where $\text{ans}(y)$ denotes the extracted option and $\text{gt}(x)$ is the ground-truth label.

Format validity reward. To stabilize multi-stage generation during RL, we enforce a strict output structure. A response is considered valid only if it contains three tag-delimited sections in the fixed order `<spatial_thinking>`, `<thinking>`, and `<answer>`, with each tag properly opened and closed. We implement the format reward as a binary indicator:

$$r_{\text{fmt}}(x, y) = \mathbb{I}[\text{ValidFormat}(y)]\,, \tag{5}$$

where $\text{ValidFormat}(y)$ returns true if and only if all required tag pairs are present and their order is strictly consistent, without missing tags, swapped order, duplicated sections, or malformed closures. This constraint discourages format violations and prevents the policy from bypassing the intended two-stage reasoning behavior by directly emitting an answer.

Length regularization reward. Finally, we add a length-based shaping term to discourage both under-generation (insufficient reasoning content) and over-generation (unnecessary verbosity). Let $\ell(y)$ denote the length of the generated response (measured in tokens). We define a length reward that is activated only when the prediction is correct and the response length falls within a preferred interval:

$$r_{\text{len}}(x, y) = \begin{cases} \omega, & \text{if } r_{\text{ans}}(x, y) = 1 \text{ and } \ell_{\min} \le \ell(y) \le \ell_{\max}, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\omega > 0$ is the shaping weight and $[\ell_{\min}, \ell_{\max}]$ specifies the target length range. In all experiments, we set $\omega = 0.2$, $\ell_{\min} = 320$, and $\ell_{\max} = 512$.

Composite reward. We combine the above components into a single reward used by GRPO:

$$r(x, y) \;=\; r_{\text{ans}}(x, y) \;+\; \lambda r_{\text{fmt}}(x, y) \;+\; r_{\text{len}}(x, y), \tag{7}$$

where $\lambda \in [0, 1]$ controls the strength of format regularization. In all experiments, we set $\lambda = 0.02$. This composite reward encourages correct answers with disciplined two-stage structure, while maintaining a reasonable reasoning length that avoids both premature truncation and overlong generations.

4 Experiments

4.1 Experimental Setup

Implementation Details. Our training follows the two-stage pipeline described earlier. In the SFT stage, we fine-tune the model using a learning rate of 1×10^{-5}. In the RL stage, we apply GRPO with a smaller learning rate of 1×10^{-6} for stable policy updates. For each training instance in RL, we sample a group of K=8 trajectories to compute group-relative advantages. Unless otherwise specified, RL training is conducted for 1500 optimization steps. On an 8×H100 machine, completing 1500 RL steps takes approximately 18 hours in our implementation.

Evaluation Settings. We evaluate our model on three multi-image, multi-view benchmarks: MMSI-Bench [44], MindCube [45], and ViewSpatial-Bench [18]. MMSI-Bench focuses on multi-image spatial reasoning that requires aligning evidence across views, such as viewpoint transformations and occlusion-aware inference. MindCube tests whether models can build a consistent mental model from limited views and perform perspective-sensitive reasoning under partial observability. ViewSpatial-Bench emphasizes viewpoint-dependent spatial localization and cross-view reference frames, highlighting generalization challenges under camera viewpoint shifts. All questions in these benchmarks are multiple-choice, and we therefore report accuracy as the primary metric. Concretely, we extract the prediction from the `<answer>` field and only accept a single option letter (e.g., A/B/C/D); outputs with format violations or non-matching parses are counted as incorrect. Our decoding and inference hyperparameters follow the recommended settings of Qwen3-VL-4B [3].

4.2 Quantitative Results

Table 1 summarizes the main results on three multi-view benchmarks. Our ViewFusion achieves the best overall performance among open-source 4B-scale models, with clear gains over the Qwen3-VL-4B family. In particular, ViewFusion (SFT+RL) improves Qwen3-VL-4B-Instruct by +5.3% on MMSI-Bench (35.4% vs. 30.1%), and yields a large improvement on MindCube (77.0% vs. 37.0%), indicating substantially stronger cross-view reasoning that benefits from

Table 1: Overall accuracy (%) on three multi-view spatial reasoning benchmarks (MMSI-Bench, MindCube, and ViewSpatial) for a range of proprietary and open-source MLLMs, including our ViewFusion variants

Model	Size	MMSI	MindCube	ViewSpatial	Avg.
RandomChoice	–	25.0	33.0	26.3	28.1
Closed-source models					
Gemini-2.5-Pro	–	38.0	57.6	46.0	47.2
GPT-5	–	41.8	56.3	45.5	47.9
Gemini-3-Pro-Preview	–	45.2	70.8	50.3	55.4
Open-source General Models					
InternVL3-2B [52]	2B	26.5	37.5	32.5	32.2
InternVL3-8B [52]	8B	28.0	41.5	38.6	36.0
Qwen2.5-VL-3B-Instruct [4]	3B	28.6	37.6	31.9	32.7
Qwen2.5-VL-7B-Instruct [4]	7B	26.8	36.0	36.8	33.2
Qwen3-VL-2B-Instruct [3]	2B	28.9	34.5	36.9	33.4
Qwen3-VL-4B-Instruct [3]	4B	30.1	37.0	42.5	36.5
Qwen3-VL-8B-Instruct [3]	8B	31.1	29.4	42.2	34.2
Spatial Intelligence Models					
SpatialLadder-3B [20]	3B	27.4	43.4	39.8	36.9
Spatial-MLLM-4B [36]	4B	26.1	33.4	34.6	31.4
SpaceR-7B [25]	7B	27.4	37.9	35.8	33.7
ViLaSR-7B [38]	7B	30.2	35.1	35.7	33.7
Cambrian-S-3B [43]	3B	25.2	32.5	39.0	32.2
Cambrian-S-7B [43]	7B	25.8	39.6	40.9	35.4
VST-3B-RL [42]	3B	32.0	36.4	45.0	37.8
VST-7B-RL [42]	7B	34.8	39.1	42.4	38.8
Ours					
ViewFusion (SFT)	4B	32.4	68.5	45.1	48.7
ViewFusion (SFT+RL)	4B	**35.4**	**77.0**	**45.4**	**52.6**

our two-stage training and GRPO alignment. On ViewSpatial, ViewFusion remains competitive at 45.4% and consistently outperforms Qwen3-VL-4B-Instruct (42.5%). Overall, these results demonstrate that explicitly optimizing for multi-view reasoning yields consistent gains across diverse evaluation settings.

To better understand the source of the improvements, Table 2 reports a fine-grained breakdown on MMSI-Bench comparing ViewFusion with Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking. Overall, ViewFusion achieves a 17.6% relative improvement over Qwen3-VL-4B-Instruct on MMSI-Bench (35.4% vs. 30.1%), suggesting that our explicit cross-view pre-alignment yields more effective multi-view reasoning rather than relying on view-local shortcuts.

Notably, ViewFusion also outperforms Qwen3-VL-4B-Thinking (35.4% vs. 29.0%), a reasoning-focused model trained with large amounts of high-quality chain-of-thought data. This comparison highlights that simply encouraging longer or higher-quality deliberation is insufficient for multi-view spatial reasoning; explicitly enforcing cross-view spatial consistency provides additional and complementary benefits.

Fig. 3: Qualitative examples on MMSI-Bench. The red boxes highlight the same visual elements observed from different viewpoints across the two images. Compared with Qwen3-VL-4B-Instruct, ViewFusion better aligns cross-view correspondences and infers the underlying viewpoint change, leading to correct answers.

4.3 Qualitative Analysis

Figure 3 presents representative MMSI-Bench examples, illustrating that multi-view spatial reasoning requires more than view-local image description. The red boxes indicate the same visual elements observed under different viewpoints. Although a strong baseline (Qwen3-VL-4B-Instruct) can often describe salient objects within each image, it frequently fails to establish cross-view spatial consistency, such as how the camera moves between views, which elements correspond under viewpoint change, and how visibility or occlusion evolves. For instance, in the first example, Qwen3-VL-4B-Instruct mainly summarizes the contents of Image 1 and Image 2 and highlights their differences, but does not reason about the spatial relationship between the two views. In contrast, ViewFusion performs explicit spatial pre-thinking before question solving. The brown `<spatial_thinking>` segment in the figure exemplifies this behavior by linking the boxed elements across views and explicitly inferring the viewpoint transformation, e.g., *"the second image seems to be taken from a position that is rotated significantly to the left"*. This cross-view alignment yields a consistent spatial interpretation that the subsequent reasoning stage can condition on, enabling correct predictions in cases where the answer depends on genuine viewpoint reasoning rather than single-image cues or purely descriptive summaries.

Table 2: Fine-grained accuracy (%) breakdown on MMSI-Bench, comparing ViewFusion with Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking across subcategories of positional relationships, attributes, motion, and MSR.

Models	Positional Relationship						Attribute		Motion		MSR	Avg.
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.	–	
Qwen3-VL-4B-Thinking	25.8	26.6	34.5	33.7	25.9	36.1	48.4	28.8	21.6	26.3	23.2	29.0
Qwen3-VL-4B-Instruct	30.1	34.0	29.6	34.9	29.4	39.8	45.3	19.7	21.6	23.7	26.7	30.1
ViewFusion	46.2	41.5	30.9	44.2	21.2	53.0	35.9	34.9	40.5	32.9	23.2	35.4

Table 3: Ablation study on MMSI-Bench (accuracy, %), analyzing the impact of free-form reasoning, removing GRPO, and removing the format reward on fine-grained subcategories and overall performance.

Models	Positional Relationship						Attribute		Motion		MSR	Avg.
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.	–	
Full Method	46.2	41.5	30.9	44.2	21.2	53.0	35.9	34.9	40.5	32.9	23.2	35.4
Free format Reasoning + RL	37.6	31.9	25.9	39.5	28.2	49.4	46.9	27.3	31.1	29.0	28.3	33.4
w/o GRPO	28.0	27.7	35.8	37.2	29.4	47.0	45.3	30.3	32.4	25.0	27.8	32.4
w/o Format Reward	40.9	30.9	33.3	46.5	32.9	51.8	29.7	40.9	37.8	34.2	22.7	35.0

4.4 Ablation Study

We conduct ablation studies on MMSI-Bench to isolate the contributions of key components, with results summarized in Table 3. Overall, the full method achieves the best average accuracy (35.4%), and the ablations reveal that our gains stem from both enforcing the two-stage protocol and leveraging GRPO-based RL for capability improvement.

First, replacing the structured two-stage reasoning with free-form reasoning under RL ("Free format Reasoning + RL") reduces the average accuracy from 35.4% to 33.4%. This drop suggests that, without an explicit spatial pre-thinking stage, the model is more likely to fall back to shortcut behaviors (e.g., premature answering from partial evidence) and fails to consistently align cross-view information before task inference.

Second, removing GRPO ("w/o GRPO") causes a larger performance degradation (35.4% $\rightarrow$ 32.4%), indicating that group-relative policy optimization plays a critical role in strengthening multi-view reasoning ability beyond what is obtained from SFT alone, likely by better matching test-time generation and directly reinforcing trajectories that lead to correct predictions.

Third, removing the format reward ("w/o Format Reward") yields only a modest decrease in average accuracy (35.4% $\rightarrow$ 35.0%). This outcome is expected because, after the SFT stage, the model already exhibits strong compliance with the required output structure, so the additional format reward in RL mainly serves as a stabilizer rather than the primary driver of accuracy gains. Taken together, the ablations confirm that the explicit two-stage reasoning supervision provides a crucial inductive bias against shortcuts, while GRPO-based RL is the main contributor to capability improvement; the format reward further helps maintain disciplined and stable generation during RL fine-tuning.

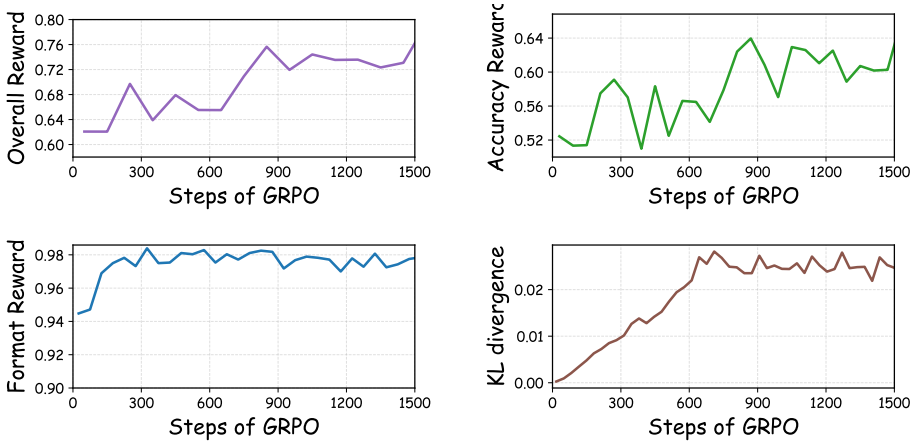

Fig. 4: GRPO training dynamics over 1500 steps. We plot the overall reward (top-left), the accuracy reward (top-right), the format reward (bottom-left), and the KL divergence to the reference policy (bottom-right).

4.5 Training Curves

Figure 4 shows the optimization dynamics of the GRPO stage over 1500 steps. The total reward increases steadily, indicating that the policy is progressively improving under the combined objective. Decomposing the reward reveals that the accuracy reward exhibits a clear upward trend, while the format reward starts at a relatively high value and quickly reaches and maintains a near-saturated level, suggesting that the desired output discipline is learned early and remains stable throughout RL training. This observation is consistent with our ablation results, where removing the format reward only causes a modest drop after SFT, implying that format compliance is largely established before RL. Meanwhile, the KL divergence increases during the initial phase and then plateaus, implying that the policy explores beyond the reference model but remains controlled under KL regularization. Overall, these curves demonstrate stable RL training that improves correctness while preserving the intended structured behavior.

5 Conclusion

We presented ViewFusion, a two-stage "think twice" framework for multi-view spatial reasoning that makes cross-view alignment an explicit first step rather than an implicit byproduct of question answering. By synthesizing structured supervision for spatial pre-thinking and further optimizing with GRPO using a correctness reward and a strict format reward, our approach mitigates shortcut behaviors that underuse available viewpoints and stabilizes multi-stage generation. Experiments on three multi-view benchmarks demonstrate consistent improvements, with particularly strong gains on MMSI-Bench and MindCube,

and fine-grained analyses show that the benefits are concentrated in categories that require genuine viewpoint reasoning. Ablations and qualitative examples further validate the contribution of each component and highlight that improved deliberation alone is insufficient without explicit cross-view spatial consistency. We hope ViewFusion serves as a simple, practical step toward more reliable multi-view reasoning in MLLMs, and motivates future work on scalable cross-view alignment objectives and broader spatial generalization.

Acknowledgments

This work is partially supported by National Key R&D Program of China under Grant No. 2024YFA1012700, by the National Natural Science Foundation of China (NSFC) under Grant No. 62402410, by Guangdong Provincial Project (No. 2023QN10X025), and by HKUST(GZ) Kunpeng&Ascend Center of Cultivation.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Batra, H., Tu, H., Chen, H., Lin, Y., Xie, C., Clark, R.: Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. arXiv preprint arXiv:2511.07403 (2025)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
7. Cheng, K., Sun, Q., Chu, Y., Xu, F., Li, Y., Zhang, J., Wu, Z.: Seeclick: Harnessing gui grounding for advanced visual gui agents. arXiv preprint arXiv:2401.10935 (2024)
8. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9062–9072 (2025)

9. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
11. Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes. arXiv preprint arXiv:2509.06266 (2025)
12. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. Guo, Z., Hong, M., Jin, T.: Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. arXiv preprint arXiv:2505.12432 (2025)
14. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14281–14290 (2024)
15. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
16. Lee, K., Lee, I., Kwak, M., Ryu, K., Hong, J., Park, J.: Spatialmosaic: A multiview vlm dataset for partial visibility. arXiv preprint arXiv:2512.23365 (2025)
17. Li, C., Zhang, C., Zhou, H., Collier, N., Korhonen, A., Vulić, I.: Topviewrs: Vision-language models as top-view spatial reasoners. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1786–1807 (2024)
18. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. arXiv preprint arXiv:2505.21500 (2025)
19. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
20. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025)
21. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
22. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
24. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
25. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)

26. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
27. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110 (2024)
28. Tao, X., Wang, Y., Cai, Y., Luo, Y., Han, K., Tang, J.: Mitigating coordinate prediction bias from positional encoding failures. arXiv preprint arXiv:2510.22102 (2025)
29. Tao, X., Wang, Y., Cai, Y., Yang, Z., Tang, J.: Understanding gui agent localization biases through logit sharpness. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 23361–23374 (2025)
30. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
31. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
32. Wang, F., Luo, N., Wu, W.: Visioncube: 3d-aware vision-language model for multi-step spatial reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3270–3279 (2025)
33. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
34. Wang, W., Zou, H., Luo, T., Huang, R., Zhao, Y., Wang, Z., Zhang, H., Qin, C., Wang, Y., Zhao, L., et al.: Video-str: Reinforcing mllms in video spatio-temporal reasoning with relation graph. arXiv preprint arXiv:2510.10976 (2025)
35. Wang, Z., Yoon, J., Yu, S., Islam, M.M., Bertasius, G., Bansal, M.: Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 28114–28128 (2025)
36. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
37. Wu, H., Huang, X., Chen, Y., Zhang, Y., Wang, Y., Xie, W.: Spatialscore: Towards comprehensive evaluation for spatial intelligence. arXiv preprint arXiv:2505.17012 (2025)
38. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)
39. Xia, J., Zang, Y., Gao, P., Li, S., Zhou, K.: Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. arXiv preprint arXiv:2505.14677 (2025)
40. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. arXiv preprint arXiv:2411.02712 (2024)
41. Yang, Q., Yao, S., Chen, W., Fu, S., Bai, D., Zhao, J., Sun, B., Yin, B., Wei, X., Zhou, J.: Humanomniv2: From understanding to omni-modal reasoning with context. arXiv preprint arXiv:2506.21277 (2025)

42. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
43. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
44. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
45. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025)
46. Yu, S., Chen, Y., Ju, H., Jia, L., Zhang, F., Huang, S., Wu, Y., Cui, R., Ran, B., Zhang, Z., et al.: How far are vlms from visual spatial intelligence? a benchmark-driven perspective. arXiv preprint arXiv:2509.18905 (2025)
47. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
48. Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.S., et al.: Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. arXiv preprint arXiv:2405.17220 **2**(3), 8 (2024)
49. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025)
50. Zhang, W., Ng, W.E., Ma, L., Wang, Y., Zhao, J., Koenecke, A., Li, B., Wanglu, W.: 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). pp. 11591–11609 (2025)
51. Zhao, R., Zhang, Z., Xu, J., Chang, J., Chen, D., Li, L., Sun, W., Wei, Z.: Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. arXiv preprint arXiv:2511.23075 (2025)
52. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)