VIEW-AID: MULTI-VIEW SPATIAL REASONING VIA AUXILIARY-IMAGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view spatial reasoning requires multimodal large language models to reconcile fragmented observations into a coherent spatial state. Direct textual chain-of-thought (CoT) construction serializes this global state into local relational statements, leaving cross-step spatial consistency implicit. Conditioning CoT generation on a reference answer further specifies the endpoint without constraining the intermediate spatial process. We introduce View-AID, which improves spatial supervision at its source by externalizing task-relevant spatial hypotheses as auxiliary images before translating them into textual CoT. This visual-first construction enriches reasoning supervision while preserving a standard inference interface that uses only the original views and question. Confidence-Guided Policy Optimization (CGPO) then complements binary outcome rewards with trace-conditioned answer confidence. On Qwen3-VL-4B, View-AID improves MindCube-Tiny and MMSI-Bench by 48.4 and 9.4 percentage points, respectively, and exceeds the strongest listed spatial baseline by 4.2 points in Overall accuracy. The gains transfer to disjoint spatial reasoning tasks and scale effectively to a larger backbone, supporting visual organization before textualization as an effective approach to multi-view spatial reasoning.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in visual recognition and language reasoning, yet integrating spatial evidence across images remains challenging (Zhao et al., 2024; Yang et al., 2025b). Multi-view spatial reasoning takes multiple images of a shared scene or object, together with a question, and requires inferring relations among objects, observers, and viewpoints by combining evidence across views (Yang et al., 2025b; Yin et al., 2025). For example, a model may need to locate an object seen in one image relative to an observer in another. Such questions require more than recognizing each image independently: view-dependent observations must be interpreted through cross-view correspondences and viewpoint transformations. Benchmarks expose persistent difficulties in these abilities, motivating methods that help models organize local observations into coherent spatial representations (Fu et al., 2024; Ma et al., 2024; Yang et al., 2025b).

Textual chain-of-thought (CoT) offers a natural interface for decomposing this reasoning into intermediate steps (Wei et al., 2022; Zhang et al., 2023). Recent approaches use capable MLLMs to construct reasoning trajectories for visual question–answer pairs: Spatial-MLLM synthesizes CoT for cold start, while ViewFusion explicitly structures cross-view pre-alignment before question answering (Wu et al., 2025a; Tao et al., 2026). Yet direct textual construction must serialize a global spatial state into local relational statements, leaving the shared configuration across viewpoints and reference objects implicit. Supplying the correct answer identifies the desired endpoint but does not specify the correspondences or transformations needed to reach it. The resulting challenge is therefore representational as well as procedural: the teacher must discover a spatial organization and verbalize it in the same step. Evidence that linguistic deliberation alone yields limited gains, together with the benefits of explicit cognitive maps, motivates introducing a spatial intermediate before verbalization (Yang et al., 2024; Yin et al., 2025; Kancheti et al., 2026).

We introduce View-AID (multi-**View** **A**uxiliary-**I**mage **D**istillation), built on the principle of organizing spatial relations visually before expressing them as textual supervision. Auxiliary Image

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

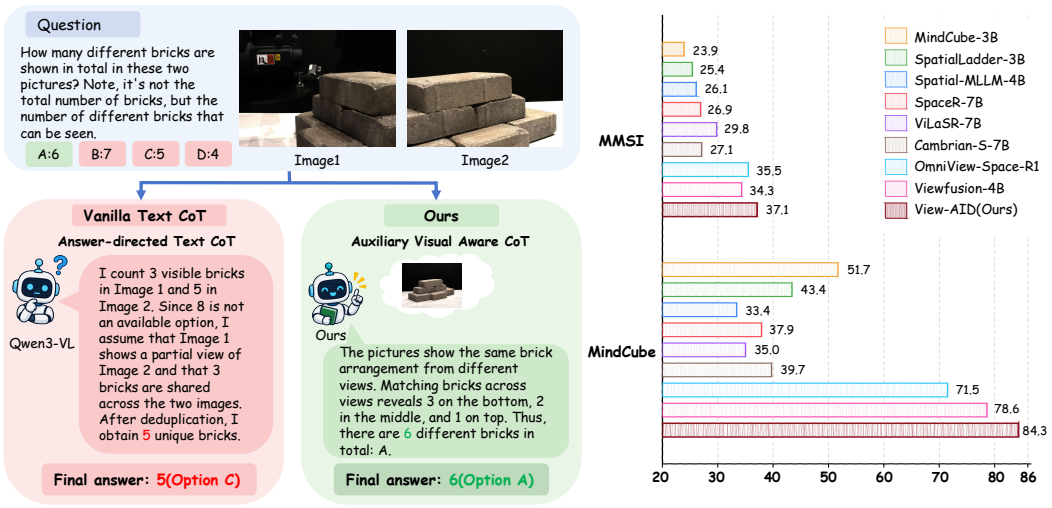


Figure 1: View-AID overview. Left: auxiliary visual representations support cross-view entity matching during CoT construction. Right: spatially specialized model accuracy on MMSI-Bench (top) and MindCube-Tiny (bottom).

Distillation (AID) asks an image generation model to externalize a task-relevant spatial hypothesis. An MLLM then interprets this hypothesis together with the original views, question, and reference answer to construct spatial CoT. Because the auxiliary image is confined to offline data construction, the target model retains the standard inference interface and requires only the original views and question. Confidence-Guided Policy Optimization (CGPO) subsequently refines the learned policy with trace-conditioned answer confidence in addition to correctness and format rewards. The two stages therefore address complementary bottlenecks: AID improves the structure of the supervision, and CGPO improves how the policy selects among the resulting reasoning behaviors.

Experiments reveal a consistent pattern across multi-view reasoning, out-of-distribution transfer, and model scales. On Qwen3-VL-4B, View-AID reaches 84.3% on MindCube-Tiny and 37.1% on MMSI-Bench, yielding the highest two-benchmark average among the compared models (Table 1). Its gains extend to SPAR-Bench and SpinBench and remain strong on single-image spatial tasks, indicating that the learned spatial organization transfers beyond the training setting. Our contributions can be summarized as follows:

- **Visual-first spatial reasoning supervision.** We identify a key limitation of direct textual CoT for multi-view spatial reasoning: serializing a global spatial state into local relational statements leaves cross-view spatial organization implicit. We therefore advocate organizing spatial structure visually before verbalizing it into reasoning supervision.

- **Auxiliary-image distillation and confidence-guided optimization.** We propose View-AID with two complementary components. AID externalizes task-relevant spatial hypotheses through generated auxiliary images and distills them into textual CoT; CGPO then refines the policy using trace-conditioned answer confidence beyond binary outcome rewards.

- **Consistent gains and broad transfer.** View-AID raises the Qwen3-VL-4B backbone by 28.9 points in Overall accuracy and surpasses the strongest listed spatial baseline by 4.2 points. The gains transfer across disjoint task formulations and scale effectively to a larger backbone, establishing visual-first supervision as a broadly effective paradigm for spatial reasoning in MLLMs.

2 Related Work

2.1 Reasoning in Multimodal Large Language Models

MLLMs integrate visual perception with language generation through multimodal conditioning and instruction tuning (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023b; Wang et al., 2024; Li et al., 2026b; 2025c). Building on CoT prompting (Wei et al., 2022), Multimodal-CoT separates rationale generation from answer inference to incorporate visual evidence into intermediate reasoning (Zhang et al., 2023). Reasoning trajectories also serve as supervision: Spatial-MLLM synthesizes textual CoT for cold start, while ViewFusion trains a structured sequence of cross-view pre-alignment and question answering (Wu et al., 2025a; Tao et al., 2026). These approaches make the construction of intermediate supervision an important part of learning multimodal reasoning.

The form of this supervision matters. Studies of language reasoning find that prompting a model to justify a supplied answer can encourage backward rationalization (Lee et al., 2026); visual spatial evaluations likewise report that longer linguistic deliberation alone does not consistently improve performance (Yang et al., 2024; Kancheti et al., 2026). Reinforcement learning offers a complementary approach, with GRPO optimizing sampled responses through group-relative rewards (Shao et al., 2024). View-AID addresses both supervision construction and subsequent optimization: auxiliary images provide visual intermediates for constructing spatial CoT, and CGPO augments outcome rewards with confidence in the answer conditioned on the generated trace.

2.2 Spatial Reasoning in Multimodal Models

Early benchmarks such as VSR and What'sUp expose weaknesses in recognizing spatial relations within individual images (Liu et al., 2023a; Kamath et al., 2023). SpatialVLM addresses this gap by deriving qualitative and quantitative spatial supervision from RGB images (Chen et al., 2024). Moving from image-level relations to scene-level understanding, 3D-LLM incorporates 3D features into language models (Hong et al., 2023), and Spatial-MLLM combines semantic and geometric encoders for spatial inference from 2D observations (Wu et al., 2025a). These approaches demonstrate the value of spatial supervision and geometric representations beyond generic visual recognition.

Multiple observations introduce the additional challenge of relating partial evidence across viewpoints. BLINK, 3DSRBench, and MMSI-Bench evaluate complementary aspects of depth, perspective, correspondence, and spatial transformation (Fu et al., 2024; Ma et al., 2024; Yang et al., 2025b). SAT and SpatialLadder develop spatial abilities through dedicated training data and curricula (Ray et al., 2024; Li et al., 2025b), while SpaceR and Cambrian-S extend spatial learning to video (Ouyang et al., 2025; Yang et al., 2025a). Explicit intermediate representations offer another route: Thinking in Space and MindCube study cognitive maps (Yang et al., 2024; Yin et al., 2025), ViLaSR interleaves reasoning with visual drawing (Wu et al., 2025b), and OmniView-Space constructs spatial maps and graphs from reconstructed geometry (Li et al., 2026a).

These findings motivate our use of visual intermediates to organize spatial relations. Our focus is their role in data construction: View-AID generates task-adaptive auxiliary images, converts their spatial content into textual supervision, and trains the target MLLM using the original views and question. This design accommodates different auxiliary image types while preserving the standard text-generation interface at inference time.

3 Method

3.1 Problem Formulation and Framework Overview

Given views $\mathcal{I} = \{I_1, \ldots, I_n\}$ and a question q, our goal is to learn a policy π_θ that generates a spatial reasoning trace z followed by an answer $\hat{y}$. Each training example also provides a reference answer y. The central design principle is to expose global spatial structure while constructing supervision, then internalize that structure so that no external tool is needed at inference time. Auxiliary Image Distillation (AID) implements the first step by externalizing a task-relevant spatial hypothesis and translating it into textual reasoning supervision. Confidence-Guided Policy Optimization (CGPO) implements the second by preferring traces that support correct and well-calibrated answers.

The two stages address complementary bottlenecks. AID improves what the policy can represent by turning a question-conditioned visual hypothesis into explicit cross-view relations, while CGPO improves how the policy selects among the resulting reasoning paths. Throughout AID, the original observations remain the evidence anchor: the auxiliary image proposes a spatial organization, but the interpreter must recover a trace that supports the reference answer from the views. CGPO then operates entirely in the target policy's response space, comparing rollouts generated for the same prompt. This separation yields a clean training-to-inference transition: external visual synthesis enriches supervision during data construction, whereas deployment retains the original input interface and incurs no auxiliary generation cost.

3.2 Auxiliary Image Distillation

AID follows a visualize–interpret–distill process (Figure 2). These operations separate tasks that direct textual CoT must solve simultaneously. Visualization proposes a global organization of the fragmented views; interpretation grounds that proposal in the observed evidence and target question; and distillation transfers the grounded relations into a policy that no longer needs the visual workspace. This decomposition makes the role of the auxiliary image explicit without changing the target model's inference interface. The construction for one example is summarized in Algorithm 1. Here, P_ω denotes the instruction proposer and Valid denotes the format and answer-support checks used to retain a trace.

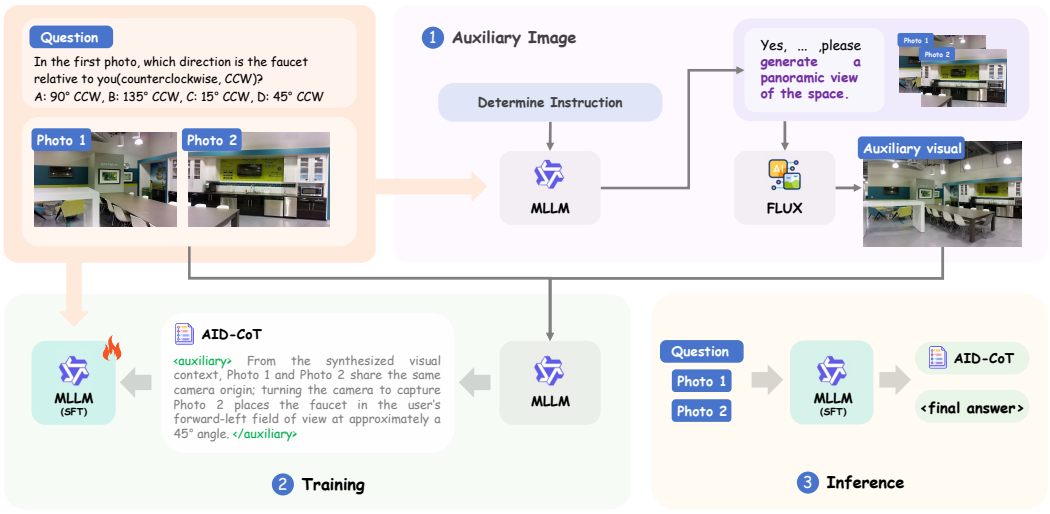


Figure 2: Auxiliary Image Distillation. A task-adaptive instruction prompts an image generation model to externalize a spatial hypothesis. An MLLM interprets this hypothesis with the original views, question, and reference answer to construct CoT supervision. The target MLLM is trained on the resulting traces and uses only the original views and question at inference time.

Generating auxiliary images. An MLLM first maps $(\mathcal{I}, q)$ to an instruction u that specifies the spatial configuration most useful for answering the question. Conditioned on the original views as visual references, an image generation model G_ϕ produces an auxiliary hypothesis $H = G_\phi(\mathcal{I}, u)$. The instruction is task-adaptive: it primarily requests viewpoint transformations, but can instead elicit other representations, such as panoramic views, when they better expose the required relations. In this way, H serves as an explicit hypothesis about how the fragmented observations fit together, rather than as an additional input required at test time.

Constructing spatial CoT. We next ask an MLLM E_ψ to interpret H against the original views, question, and reference answer: $z = E_\psi(\mathcal{I}, q, y, H)$. The construction prompt requires the trace to identify cross-view object correspondences, explain relevant viewpoint changes, and derive the relative positions that support y. This step converts the holistic spatial hypothesis into an auditable

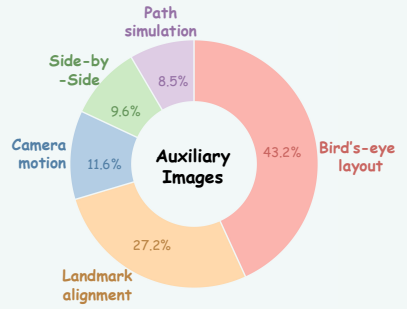


Figure 3: Auxiliary-image category distribution in the AID data. Bird's-eye layouts and landmark alignment are the two dominant construction patterns.

Algorithm 1. Auxiliary Image Distillation

Require: $\mathcal{D} = \{(\mathcal{I}, q, y)\}$, G_ϕ, E_ψ, π_θ
Ensure: Spatially distilled policy π_θ

$\mathcal{D}_{\text{AID}} \leftarrow \varnothing$
for all $(\mathcal{I}, q, y) \in \mathcal{D}$ **do**
  $u \leftarrow P_\omega(\mathcal{I}, q)$ ▷ view instruction
  $H \leftarrow G_\phi(\mathcal{I}, u)$ ▷ spatial hypothesis
  $z \leftarrow E_\psi(\mathcal{I}, q, y, H)$ ▷ reasoning trace
  if $\text{Valid}(z, y)$ **then**
    $r \leftarrow [\texttt{auxiliary} : z;\ \texttt{answer} : y]$
    $\mathcal{D}_{\text{AID}} \leftarrow \mathcal{D}_{\text{AID}} \cup \{(\mathcal{I}, q, r)\}$
  end if
end for
$\theta \leftarrow \arg\min_\theta \mathcal{L}_{\text{distill}}(\theta; \mathcal{D}_{\text{AID}})$
Discard G_ϕ and E_ψ after distillation
Inference: condition only on $(\mathcal{I}, q)$
return π_θ

sequence of textual relations. We retain traces that pass format and answer-support checks, so the auxiliary image influences training only through the distilled reasoning trace.

Supervised fine-tuning. Finally, we internalize the distilled spatial reasoning in the target policy. We construct response r by placing the retained trace z between `<auxiliary>` and `</auxiliary>`, followed by y between `<answer>` and `</answer>`. The resulting dataset $\mathcal{D}_{\text{AID}}$ contains triples $(\mathcal{I}, q, r)$, and we minimize the autoregressive objective

$$\mathcal{L}_{\text{distill}}(\theta) = -\mathbb{E}_{(\mathcal{I},q,r)\sim\mathcal{D}_{\text{AID}}}\left[\sum_{t=1}^{|r|} \log \pi_\theta(r_t \mid r_{<t}, \mathcal{I}, q)\right]. \tag{1}$$

Because H is used only to construct z, inference requires no image generation: the learned policy produces both the reasoning and answer blocks directly from $(\mathcal{I}, q)$.

3.3 Confidence-Guided Policy Optimization

SFT teaches the policy to imitate retained traces but does not distinguish among alternative rollouts that depart from those demonstrations. CGPO starts from the AID-SFT checkpoint and optimizes sampled responses using correctness, format, and confidence rewards. Its confidence term supplies a graded preference within the coarse groups induced by binary correctness, favoring traces that make correct answers more probable and incorrect answers less probable.

Confidence-guidance reward. For a sampled response containing trace z_i and answer $\hat{y}_i$, we define confidence as the likelihood that the current policy assigns to the generated answer tokens after first producing the corresponding reasoning trace:

$$p_i = \pi_\theta(\hat{y}_i \mid \mathcal{I}, q, z_i) = \prod_{t=1}^{|\hat{y}_i|} \pi_\theta(\hat{y}_{i,t} \mid \mathcal{I}, q, z_i, \hat{y}_{i,<t}). \tag{2}$$

Response delimiters are included in the conditioning context but omitted from the notation. Let $c_i = \mathbf{1}[\hat{y}_i = y]$ denote correctness. For a valid reasoning–answer pair, we assign the negative binary Brier-style reward (Brier, 1950) $R^{(i)}_{\text{conf}} = -(p_i - c_i)^2$. This reward favors high confidence for correct answers and low confidence for incorrect ones, coupling the quality of a sampled trace to the answer distribution it induces.

Combined reward and policy update. We combine this signal with task and format rewards:

$$R_i = R^{(i)}_{\text{acc}} + \lambda_{\text{fmt}} R^{(i)}_{\text{fmt}} + \lambda_{\text{conf}} R^{(i)}_{\text{conf}}, \tag{3}$$

where $R^{(i)}_{\text{acc}} = c_i$, $R^{(i)}_{\text{fmt}}$ evaluates the required reasoning–answer format, and $\lambda_{\text{fmt}}, \lambda_{\text{conf}} \geq 0$ control their relative weights. Correctness anchors optimization to task success, while confidence resolves preferences among responses that share the same binary outcome.

We optimize the policy using GRPO (Shao et al., 2024). For each prompt, the rollout policy samples K complete responses. Their combined rewards define the group-relative advantages

$$\bar{R} = \frac{1}{K}\sum_{j=1}^{K} R_j, \qquad A_i = \frac{R_i - \bar{R}}{\text{std}(R_1, \ldots, R_K) + \epsilon}, \tag{4}$$

where $\epsilon > 0$ stabilizes normalization. We use A_i in the clipped GRPO objective with reference-policy regularization (Shao et al., 2024). Setting $\lambda_{\text{conf}} = 0$ removes confidence guidance and recovers the vanilla GRPO with correctness and format rewards, enabling a controlled comparison of the proposed confidence signal.

4 Experiments

Table 1: Main benchmark accuracy (%). Overall averages the two *All* scores; bold marks column maxima among open models and underlining marks the second-highest open-model score. Proprietary-model scores are shown in light gray for reference only. Delta rows report gains over Qwen3-VL-4B-Instruct.

Model	MindCube-Tiny				MMSI-Bench					Overall
	Aro.	Rot.	Amo.	All	Pos.	Att.	Mot.	MSR	All	
Proprietary Models (reference only)										
GPT-4.1	59.6	49.5	47.2	50.6	29.3	33.8	38.7	31.3	31.7	41.2
GPT-5.2	69.2	95.5	41.5	58.4	42.9	47.6	36.0	40.4	42.0	50.2
Gemini-3-Pro	51.6	54.5	46.0	49.0	42.9	50.0	38.0	40.9	42.7	45.9
Open-source General Models										
InternVL3-8B	53.6	36.5	38.1	41.5	30.8	23.8	24.6	14.6	25.7	33.6
Qwen2.5-VL-7B-Instruct	21.4	38.8	29.5	29.3	28.4	21.5	29.3	<u>25.8</u>	27.1	28.2
Qwen3-VL-4B-Instruct	41.6	34.1	20.0	35.9	34.2	32.3	26.0	23.7	27.7	31.8
Qwen3-VL-8B-Instruct	52.0	45.0	34.8	40.9	32.6	30.0	24.0	21.2	29.0	34.9
Spatial Intelligence Models										
MindCube-3B	67.6	34.0	51.0	51.7	23.9	24.6	18.7	25.3	23.9	37.8
SpatialLadder-3B	50.8	35.0	43.2	43.4	30.3	23.3	16.0	21.2	25.4	34.4
Spatial-MLLM-4B	36.0	39.0	30.5	33.4	28.5	25.4	18.0	**26.3**	26.1	29.8
SpaceR-7B	49.2	35.0	34.2	37.9	29.1	29.4	21.9	22.5	26.9	32.4
ViLaSR-7B	44.4	35.5	31.0	35.0	35.9	26.0	21.0	23.2	29.8	32.4
Cambrian-S-7B	50.8	37.0	35.9	39.7	-	-	-	-	27.1	33.4
OmniView-Space-R1	68.0	53.0	79.2	71.5	41.2	<u>35.4</u>	28.7	<u>25.8</u>	<u>35.5</u>	53.5
ViewFusion-4B	70.8	**89.5**	78.2	78.6	38.1	32.3	**34.0**	<u>25.8</u>	34.3	56.5
Ours										
View-AID-SFT	<u>82.8</u>	<u>77.5</u>	<u>86.0</u>	<u>83.6</u>	<u>41.8</u>	**36.0**	28.7	22.2	<u>35.5</u>	<u>59.6</u>
Δ *SFT–Base*	+41.2	+43.4	+66.0	+47.7	+7.6	+3.7	+2.7	-1.5	+7.8	+27.8
View-AID-CGPO	**85.2**	74.0	**87.3**	**84.3**	**43.6**	**36.0**	<u>31.4</u>	23.7	**37.1**	**60.7**
Δ *CGPO–Base*	+43.6	+39.9	+67.3	+48.4	+9.4	+3.7	+5.4	+0.0	+9.4	+28.9

4.1 Experimental Setting

We organize the evaluation around three questions: whether visual-first supervision improves the primary multi-view benchmarks, whether the learned spatial operations transfer beyond the training distribution, and how AID and confidence guidance contribute to the final behavior.

Tasks and data. We use MindCube-Tiny (Yin et al., 2025) and MMSI-Bench (Yang et al., 2025b) as the primary multi-view benchmarks and report multiple-choice accuracy. AID uses the ViewFu-

sion SFT source data, whereas CGPO uses its RL data (Tao et al., 2026). We define *Overall* as the unweighted mean of the two benchmark-level scores.

Implementation details. Unless stated otherwise, Qwen3-VL-4B-Instruct is the target MLLM. Flux.2-klein-4B generates the auxiliary images, and Qwen3-VL constructs and filters the textual traces. We train on the resulting AID data for one epoch with a learning rate of 1×10^{-5}, then initialize CGPO from this checkpoint and optimize with a learning rate of 1×10^{-6}. All runs use bfloat16. To isolate the effect of the training objective, controlled strategy comparisons share image preprocessing, evaluation prompts, data splits, and decoding protocol.

4.2 Main Results and Generalization

Baselines. Table 1 compares View-AID with scores reported in prior publications, while the controlled training-strategy ablation in Table 5 isolates the contribution of each component. We compare with proprietary models GPT-4.1 (OpenAI, 2025a), GPT-5.2 (OpenAI, 2025b), and Gemini-3-Pro (Google DeepMind, 2025); open-source generalist models from the InternVL3 (Zhu et al., 2025), Qwen2.5-VL (Bai et al., 2025), and Qwen3-VL (Qwen Team, 2025) families; and spatially specialized models including MindCube (Yin et al., 2025), SpatialLadder (Li et al., 2025b), Spatial-MLLM (Wu et al., 2025a), SpaceR (Ouyang et al., 2025), ViLaSR (Wu et al., 2025b), Cambrian-S (Yang et al., 2025a), OmniView-Space (Li et al., 2026a), and ViewFusion (Tao et al., 2026).

Main benchmark results. Table 1 shows that visual-first supervision transforms the general-purpose backbone into a strong spatial reasoner. The gain is broad across MMSI-Bench and particularly pronounced on MindCube-Tiny, where reconstructing a coherent configuration across views is central. CGPO preserves the large SFT improvement while strengthening position- and motion-sensitive reasoning. The complete method reaches the highest Overall score among open models (60.7), 4.2 points above the strongest listed spatial baseline, and leads the MindCube-Tiny aggregate as well as arrangement, amount, and position reasoning.

Extension to Qwen3-VL-8B. Increasing backbone capacity preserves the characteristic AID improvement pattern (Table 2): gains remain broad and are strongest on arrangement and amount reasoning, the categories that most directly benefit from an explicit global organization. Overall rises by 16.8 points, indicating that visual-first supervision scales effectively in the Qwen3-VL family.

Table 2: AID-based SFT on Qwen3-VL-8B. Accuracy (%); Overall averages the two *All* scores. The delta row reports gains over Qwen3-VL-8B.

Model	MindCube-Tiny				MMSI-Bench					Overall
	Aro.	Rot.	Amo.	All	Pos.	Att.	Mot.	MSR	All	
Qwen3-VL-8B	52.0	45.0	34.8	40.9	32.6	**30.0**	24.0	21.2	29.0	34.9
View-AID-SFT (8B)	**78.4**	**49.0**	**71.0**	**68.6**	**42.2**	**30.0**	**28.0**	**21.8**	**34.9**	**51.7**
Δ *SFT–Base*	+26.4	+4.0	+36.2	+27.7	+9.6	+0.0	+4.0	+0.6	+5.9	+16.8

Table 3: Out-of-distribution benchmark accuracy (%). Deltas are measured relative to Base.

Model	SPAR-Bench			SpinBench		
	OSR	SI	DP	MR	SRT	RC
Base	46.8	34.0	34.1	24.3	41.4	41.1
SFT	56.5	38.4	34.8	**36.7**	**56.2**	44.5
Δ *SFT*	+9.7	+4.4	+0.7	+12.4	+14.8	+3.4
CGPO	**58.8**	**39.5**	**37.3**	35.3	50.0	**47.9**
Δ *CGPO*	+12.0	+5.5	+3.2	+11.0	+8.6	+6.8

Table 4: Single-image benchmark accuracy (%). Deltas are measured relative to Base.

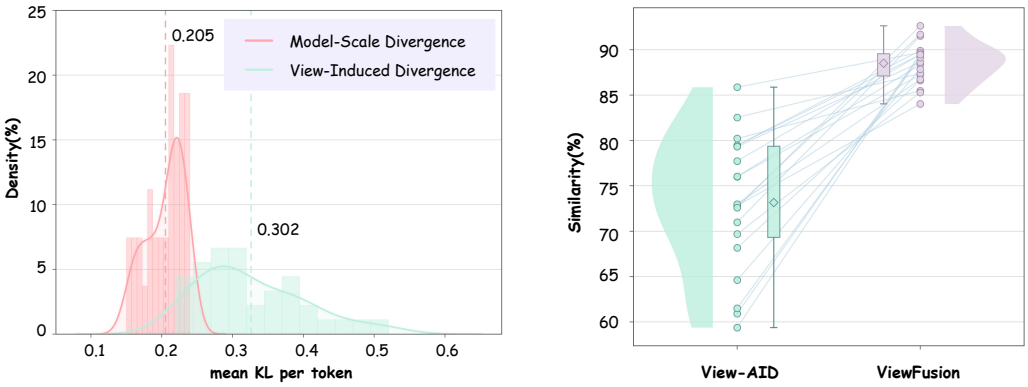
Model	SPAR-Bench				ViewSpatial		
	SI	DP-OO	DP-OC	C-OVO	C-RD	P-OVO	P-RD
Base	28.8	19.7	28.3	**25.7**	26.4	23.8	27.6
SFT	**38.7**	**26.8**	**33.2**	23.4	47.2	**77.0**	35.9
Δ *SFT*	+9.9	+7.1	+5.0	-2.3	+20.8	+53.2	+8.3
CGPO	37.9	25.5	32.4	25.6	**47.6**	73.3	**37.2**
Δ *CGPO*	+9.1	+5.8	+4.2	-0.1	+21.2	+49.5	+9.6

Generalization and retention of spatial capabilities. Tables 3 and 4 demonstrate that the learned spatial capabilities transfer beyond the training setting, covering both disjoint multi-view and spatial-transformation tasks as well as single-view evaluation. Almost all categories improve over Base

under both trained variants, with AID-SFT showing particular strength in mental rotation, spatial-relation transformation, and person-perspective orientation, while CGPO performs best on SPAR-Bench tasks, rotation classification, and camera- and person-relative direction. The largest gains on viewpoint-sensitive categories further suggest that the learned representation captures reusable transformations across observers and reference frames.

4.3 Diagnostic Analysis

Having established performance and transfer, we sample a compact diagnostic subset from the benchmark data for controlled behavioral analysis. On this shared subset, we examine two signatures of the proposed pipeline: the sensitivity of the response distribution to organized visual context and the breadth of the reasoning traces explored by the learned policy (Figure 4).



(a) Response-distribution divergence.

(b) Similarity across reasoning rollouts.

Figure 4: Dataset-level diagnostics of visual context and reasoning diversity on a sampled benchmark subset. (a) Distributions of per-response mean token-level KL relative to Qwen3-VL-4B with two views, pooled across the sampled examples: blue increases the model to 8B, whereas red supplies the 4B model with an additional global view. (b) Mean pairwise cosine similarity of reasoning embeddings on paired MMSI-Bench questions. Green denotes View-AID and purple denotes ViewFusion; lines connect the same question and diamonds mark means. Lower similarity indicates more diverse sampled traces.

Visual context changes the response distribution. Across the sampled diagnostic subset, adding a global view shifts the response distribution more than increasing model size from 4B to 8B (Figure 4(a)). This contrast illustrates the leverage of organized visual context: changing the available spatial structure can affect generation more strongly than adding parameters alone.

Diversity of spatial reasoning rollouts. Across the sampled MMSI-Bench diagnostic subset, View-AID produces more diverse reasoning traces than ViewFusion under matched decoding conditions (Figure 4(b)). The lower mean similarity indicates that the policy explores a broader set of spatial hypotheses instead of repeatedly following near-identical reasoning paths. Appendix D provides the full protocol.

4.4 Component Analysis

Effect of training strategy. We return to the two stages of the method and isolate their contributions under a shared Qwen3-VL-4B-Instruct backbone. In Table 5, SFT uses the AID-constructed traces, whereas GRPO *from scratch* applies RL directly to the backbone. Starting from AID-SFT, $\lambda_{\text{conf}} = 0$ recovers standard GRPO and $\lambda_{\text{conf}} = 1$ gives the default CGPO configuration, directly isolating confidence guidance.

AID-SFT provides the decisive performance gain, lifting both benchmarks and raising Overall by 27.8 points (Table 5). In contrast, GRPO from the backbone yields only a modest aggregate improvement. The comparison identifies structured visual supervision as the foundation of the learned spatial ability, rather than policy optimization alone.

Table 5: Training-strategy and confidence-weight ablations. Accuracy (%); Overall averages the two benchmark-level *All* scores. In the lower block, $\lambda_{\text{conf}} = 0$ is standard GRPO and $\lambda_{\text{conf}} = 1$ is the default CGPO setting; bold marks column maxima.

Setting	MindCube-Tiny				MMSI-Bench					Overall
	Aro.	Rot.	Amo.	All	Pos.	Att.	Mot.	MSR	All	
Initialization and RL baseline										
Qwen3-VL-4B-Instruct	41.6	34.1	20.0	35.9	34.2	32.3	26.0	23.7	27.7	31.8
+ GRPO (from scratch)	47.2	33.0	39.3	40.0	28.6	30.2	25.3	24.2	27.4	33.7
+ AID-SFT	82.8	**77.5**	86.0	83.6	41.8	36.0	28.7	22.2	35.5	59.6
Policy optimization from AID-SFT										
+ GRPO ($\lambda_{\text{conf}} = 0$)	84.4	73.5	86.3	83.4	39.3	35.3	28.1	22.7	34.0	58.7
+ CGPO ($\lambda_{\text{conf}} = 0.1$)	84.4	75.0	**87.5**	84.4	43.0	35.2	30.7	22.7	36.4	60.4
+ CGPO ($\lambda_{\text{conf}} = 0.5$)	**85.2**	75.5	87.2	**84.5**	42.5	35.2	30.7	**24.8**	36.5	60.5
+ CGPO ($\lambda_{\text{conf}} = 1$)	**85.2**	74.0	87.3	84.3	**43.6**	36.0	31.4	23.7	37.1	**60.7**
+ CGPO ($\lambda_{\text{conf}} = 2$)	84.0	74.0	86.3	83.4	43.4	**36.1**	**32.0**	24.2	**37.2**	60.3

From the same SFT checkpoint, CGPO improves all seven subtasks over standard GRPO and reaches the best Overall score. The strongest refinements appear in position and motion reasoning, showing that trace-conditioned confidence adds a consistent preference signal after the core ability has been established by AID.

Sensitivity to the confidence-guidance weight. The confidence-weight sweep shows that the benefit is not tied to a single coefficient (Table 5). Every nonzero setting improves Overall over standard GRPO, and the default setting provides the strongest aggregate result while improving every subtask. Confidence guidance therefore acts as a robust refinement signal across a useful range of weights.

Role of the image-generation instruction. Task-adaptive image instructions improve every MindCube-Tiny category and strengthen position and attribute reasoning on MMSI-Bench (Figure 5). The contrast shows that image generation alone is insufficient: specifying the spatial operation needed by each question turns the auxiliary image into targeted reasoning supervision.

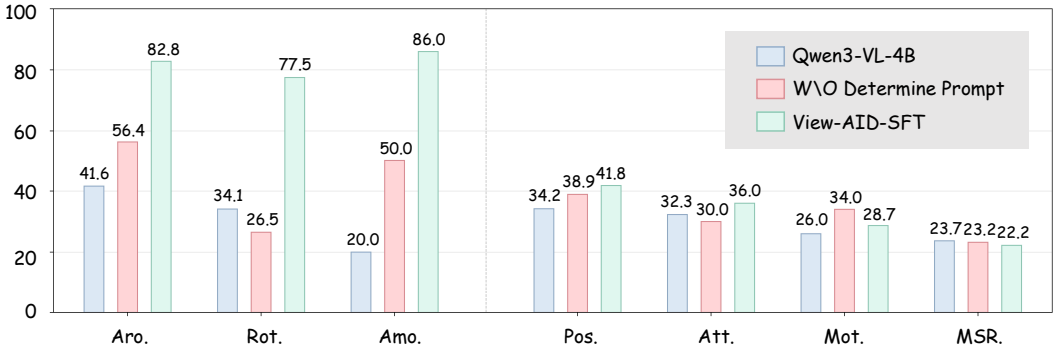


Figure 5: Image-generation instruction ablation. Accuracy (%) for the backbone, the variant without the instruction-determination prompt, and View-AID-SFT; subtask order follows Table 5.

5 Discussion

The results support a two-stage account of spatial learning: AID converts holistic visual hypotheses into reusable reasoning traces, while CGPO refines how the policy selects and calibrates those traces. This division explains the dominant gains from structured supervision and the complementary improvements from confidence guidance. Together with consistent transfer and effective scaling, these findings suggest that visual organization can serve as a powerful training-time workspace whose structure is distilled into an efficient policy model that requires only the original views at inference.

AI USE DISCLOSURE

The authors used OpenAI's ChatGPT and Codex for language editing and structural revision across the manuscript, including the Method, Experiments, Discussion, Appendix, figure captions, and LaTeX table formatting, and to summarize high-level patterns already present in author-provided results. These tools were not used to generate experimental data, conduct experiments, originate the proposed method, retrieve literature, or prove mathematical claims. The authors checked all AI-assisted text against the underlying methods and results and take full responsibility for the final manuscript.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 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. URL https://arxiv.org/abs/2502.13923.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. URL https://arxiv.org/abs/2404.12390.

Google DeepMind. Gemini 3 pro model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-pro/. Model released November 2025; model card updated May 2026.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 2023.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. *arXiv preprint arXiv:2604.16060*, 2026. URL https://arxiv.org/abs/2604.16060.

Jungseob Lee, Seungyoon Lee, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, and Heuiseok Lim. Answer-conditioned chains of thought degrade verifiable-reasoning distillation in large language models. *arXiv preprint arXiv:2607.14552*, 2026. URL https://arxiv.org/abs/2607.14552.

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. *arXiv preprint arXiv:2505.21500*, 2025a. URL https://arxiv.org/abs/2505.21500.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025b. URL https://arxiv.org/abs/2510.08531.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023.

Xudong Li, Mengdan Zhang, Peixian Chen, Jiaxi Tan, Zihao Huang, Jingyuan Zheng, Yan Zhang, Xiawu Zheng, Xing Sun, and Rongrong Ji. OmniView-Space: Reinforcing spatial reasoning via multi-perspective spatial mapping. *arXiv preprint arXiv:2607.00881*, 2026a. URL https://arxiv.org/abs/2607.00881.

Yuxuan Li, Yicheng Zhang, Wenhao Tang, Yimian Dai, Ming-Ming Cheng, Xiang Li, and Jian Yang. Visual instruction pretraining for domain-specific foundation models. *arXiv preprint arXiv:2509.17562*, 2025c.

Yuxuan Li, Yuming Chen, Yunheng Li, Ming-Ming Cheng, Xiang Li, and Jian Yang. Unifying heterogeneous multi-modal remote sensing detection via language-pivoted pretraining. In *ICML*, 2026b.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023b.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan Yuille, and Jieneng Chen. 3DSR-Bench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. URL https://arxiv.org/abs/2412.07825.

OpenAI. Introducing GPT-4.1 in the API. OpenAI Technical Report, 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. Update to GPT-5 system card: GPT-5.2. OpenAI System Card, 2025b. URL https://openai.com/index/gpt-5-system-card-update-gpt-5-2/.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. URL https://arxiv.org/abs/2504.01805.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xingjian Tao, Yiwei Wang, Yujun Cai, Yifan Song, and Jing Tang. ViewFusion: Structured spatial thinking chains for multi-view reasoning. *arXiv preprint arXiv:2603.06024*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

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

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

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. *arXiv preprint arXiv:2412.14171*, 2024. URL https://arxiv.org/abs/2412.14171.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Sain-

ing Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025a. URL https://arxiv.org/abs/2511.04670.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025b.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. URL https://arxiv.org/abs/2506.21458.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv preprint arXiv:2503.22976*, 2025a. URL https://arxiv.org/abs/2503.22976.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b. URL https://arxiv.org/abs/2506.05176.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. *arXiv preprint arXiv:2509.25390*, 2025c. URL https://arxiv.org/abs/2509.25390.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Benchmark and Evaluation Details

Task abbreviations. Table 3 reports the observer-centered multi-view (OC-MV) tasks from SPAR-Bench: OSR denotes object spatial relation, SI spatial imagination, and DP depth prediction. For SpinBench, MR denotes mental rotation, SRT spatial-relation transformation, and RC rotation classification.

Table 4 reports single-image tasks. For SPAR-Bench, SI denotes spatial imagination and DP depth prediction; OC denotes observer-centered and OO object–object. For ViewSpatial-Bench, C and P denote camera and person perspectives, respectively; OVO denotes object view orientation and RD relative direction.

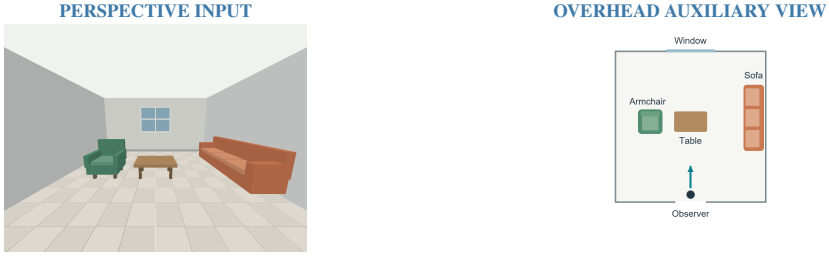
OOD and single-view evaluation. We test out-of-distribution (OOD) transfer on selected SPAR-Bench (Zhang et al., 2025a) and SpinBench (Zhang et al., 2025c) tasks whose samples and source scenes do not overlap with the SFT or RL data. SPAR-Bench evaluates observer-centered relations, counterfactual viewpoint reasoning, and depth prediction; SpinBench evaluates mental rotation (MR), spatial-relation transformation (SRT), and paired-view rotation classification (RC). We further evaluate single-image tasks from SPAR-Bench and ViewSpatial-Bench (Li et al., 2025a) to examine whether multi-view training induces spatial abilities that remain useful from camera and person perspectives. In Tables 3 and 4, Base denotes the backbone, SFT the AID-trained checkpoint, and CGPO the complete model.

B AID Data Construction and Auxiliary-Image Taxonomy

Training-example organization. Each AID example starts from benchmark views $\mathcal{I}$, question q, and reference answer y. The instruction proposer identifies the required spatial operation, the image generator externalizes it as an auxiliary view H, and the explainer converts H into a textual trace z. Retained responses place z in the `<auxiliary>` block and y in the `<answer>` block. The auxiliary image is therefore used during data construction, while inference requires only $(\mathcal{I}, q)$.

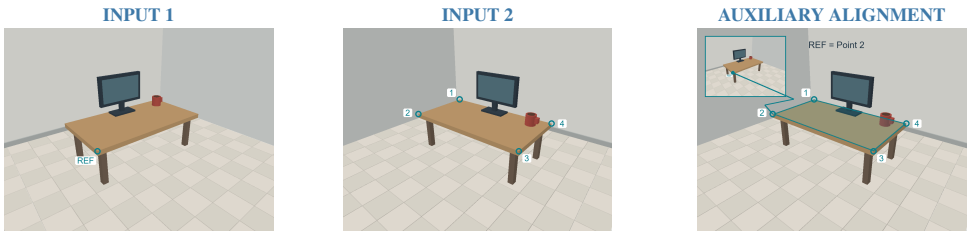
Task-adaptive visual operations. Figures 6 and 7 organize five recurring operations: bird's-eye layout, landmark alignment, camera-motion schematic, side-by-side comparison, and path simulation. They respectively expose a common coordinate frame, cross-view identity, camera displacement, shared visual evidence, or an instructed action path. The synthetic 3-D scenes illustrate these construction choices; Appendix C presents generated auxiliary images and distilled traces on MMSI-Bench and MindCube-Tiny.

(a) Bird's-eye layout reconstruction



Question. Facing the window from the doorway, is the green armchair to the left or right of the coffee table? **Construction.** Place the window, table, sofa, and armchair in a common overhead layout; mark the observer and forward direction to expose observer-relative left/right.

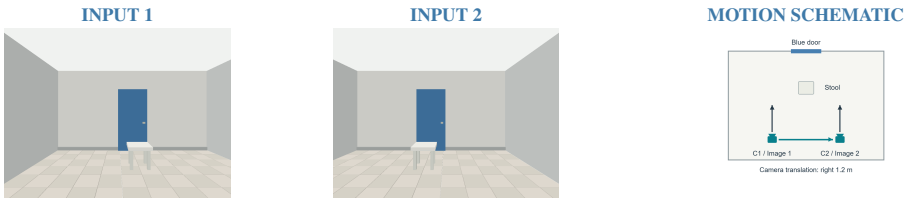
(b) Cross-view landmark alignment



Question. Which point in Input 2 corresponds to the desk corner marked REF in Input 1? **Construction.** Align the tabletop boundaries, transfer the reference point, and connect it to Point 2 in the second view.

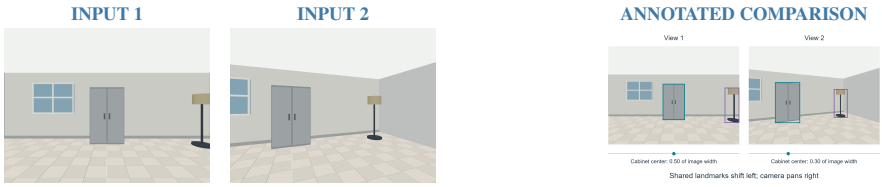
Figure 6: Illustrative auxiliary-image construction patterns (I). Bird's-eye reconstruction externalizes an observer-centered layout, while landmark alignment makes cross-view correspondence explicit. The scenes are deterministic synthetic illustrations of the visual operations.

(c) Camera-motion schematic



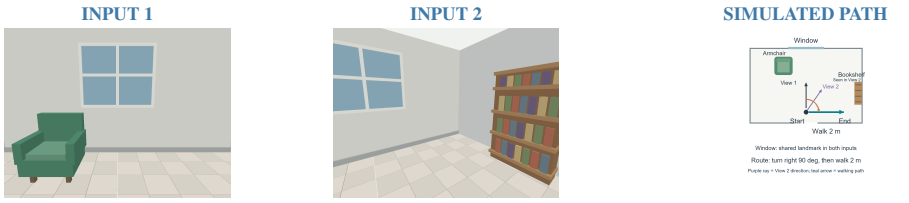
Question. With viewing direction fixed, did the camera move left or right? **Construction.** Place both camera poses relative to the shared door and stool, retain parallel viewing arrows, and draw the lateral translation between them.

(d) Side-by-side view comparison



Question. At a fixed position, does the camera pan left or right? **Construction.** Preserve both views, mark the same cabinet and lamp, and use their leftward image shift to reveal a rightward camera pan.

(e) Turn-and-move path simulation



Question. After a right turn and a two-meter walk, which object is approached? **Construction.** Match the shared window, integrate both views in an overhead layout, and draw the instructed clockwise turn and eastward path toward the bookshelf.

Figure 7: Illustrative auxiliary-image construction patterns (II). Camera-pose diagrams, annotated view comparisons, and action-path simulations expose the transformations that connect partial observations to the queried spatial relation.

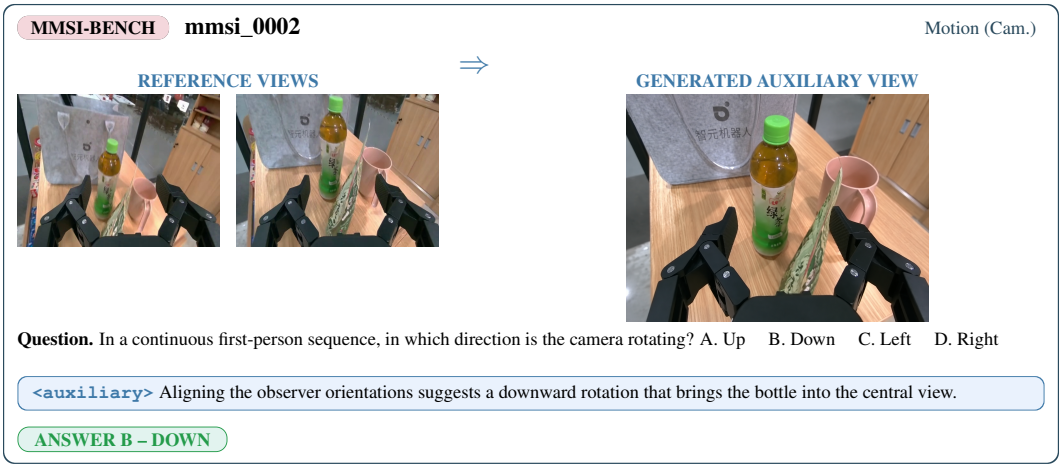
C Qualitative AID Examples

We present six retained examples spanning the two main benchmarks. Each card follows the same data flow: original views, generated auxiliary view, distilled `<auxiliary>` trace, and final answer. Together, the cases cover camera motion, multi-view spatial relations, attribute comparison, action-conditioned navigation, mental rotation, and observer translation.

The cards below retain the question-conditioned evidence flow and summarize the spatial transformation that supports each answer.

Table 6: Coverage of the qualitative AID examples and their principal spatial operations.

Dataset	Example	Task	Principal spatial operation	Answer
MMSI-Bench	mmsi_0002	Motion	Align object evidence to infer downward camera rotation	B
MMSI-Bench	mmsi_0012	MSR	Place the bed, desk, and sink in an observer-centered layout	C
MMSI-Bench	mmsi_0007	Attribute	Align views before comparing object extent	A
MindCube-Tiny	mindcube_0001	Among	Compose a right turn with forward movement	A
MindCube-Tiny	mindcube_0605	Rotation	Rotate the reference frame and update left/right relations	C
MindCube-Tiny	mindcube_0803	Around	Reconcile landmarks to infer diagonal observer translation	A



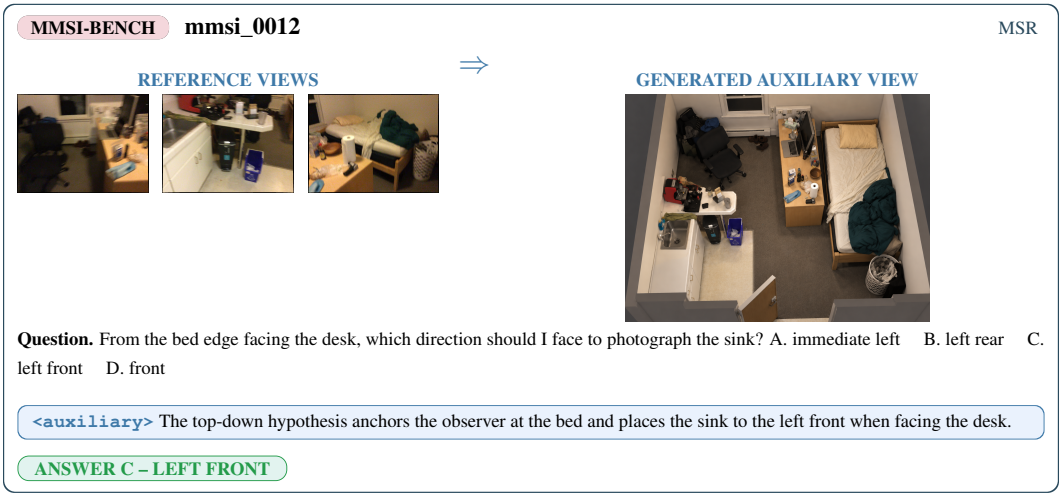
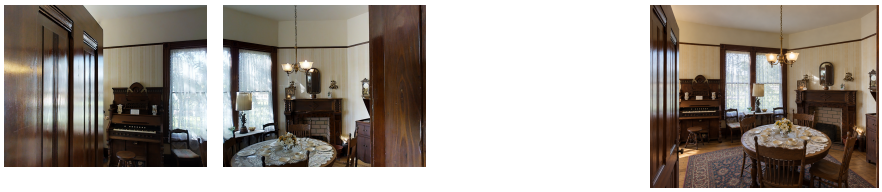


Figure 8: MMSI-Bench reasoning examples. Cross-view alignment supports the downward camera-rotation inference, while an integrated overhead hypothesis resolves the sink direction from the bed-facing-desk reference frame.

MMSI-BENCH **mmsi_0007** Attribute (Meas.)

REFERENCE VIEWS ⇒ GENERATED AUXILIARY VIEW



Question. Which is longer: the piano or the table under the mirror? A. table B. depends C. same length D. piano

<auxiliary> Aligning the observer orientations makes the piano and table directly comparable and indicates that the table under the mirror is longer.

ANSWER A – TABLE

MINDCUBE-TINY **mindcube_0001** Among

REFERENCE VIEWS ⇒ GENERATED AUXILIARY VIEW



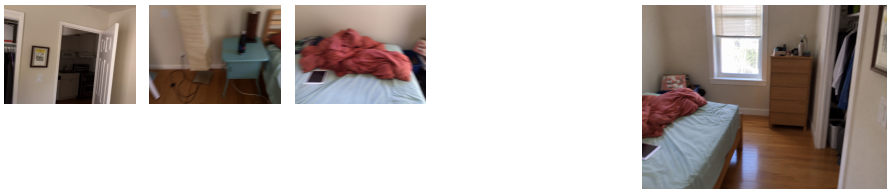
Question. From image 1, if I turn right and move forward, will I get closer to the pink plush toy and headboard? A. Yes B. No

<auxiliary> Aligning the initial orientation with image 1 and composing a right turn with forward movement places the observer in front of the plush toy and headboard.

ANSWER A – YES

MINDCUBE-TINY **mindcube_0605** Rotation

REFERENCE VIEWS ⇒ GENERATED AUXILIARY VIEW



Question. From image 3, after a 90-degree right turn, what is to my left? A. bedside table B. door C. bed and red blanket

<auxiliary> Align image 3 with the auxiliary view, rotate the observer frame 90 degrees clockwise, and update the left/right relation; the bed and red blanket fall to the left.

ANSWER C – BED AND RED BLANKET

Figure 9: Qualitative reasoning examples across attribute comparison, action-conditioned navigation, and observer-frame rotation. Each card connects the original views, generated auxiliary view, distilled trace, and grounded answer.

MINDCUBE-TINY **mindcube_0803** Around

REFERENCE VIEWS ⇒ GENERATED AUXILIARY VIEW



Question. From the first view to the second, in which direction did I move? A. diagonal forward-left B. diagonal forward-right C. left D. right

<auxiliary> Aligning the kiosk and traffic cone across views reveals a diagonal forward-left observer translation.

ANSWER A – DIAGONAL FORWARD-LEFT

Figure 10: Landmark-based observer-motion reasoning on MindCube-Tiny. Aligning the kiosk and traffic cone across views makes the diagonal forward-left translation explicit.

D Diagnostic Analysis Protocols

We construct a compact diagnostic subset by sampling examples from the benchmark data. All compared conditions use the same selected examples, and each probe shares its decoding and scoring procedures across conditions. This paired design focuses the analysis on behavioral differences induced by visual context or learned policy rather than variation in the evaluated inputs.

Response-distribution sensitivity to visual context. For every selected example in the response-distribution probe, we compare two interventions against a common reference condition, Qwen3-VL-4B with the original two views. The first uses Qwen3-VL-8B with the same views; the second retains the 4B model and supplies an additional global view, together with an instruction identifying its role. Responses are sampled under each intervention across the subset, and both the intervention and reference conditions are evaluated on identical response prefixes. For each response r, we measure

$$\overline{D}_{\mathrm{KL}}(r) = \frac{1}{|r|} \sum_{t=1}^{|r|} D_{\mathrm{KL}}(P(\cdot \mid r_{<t}) \,\|\, Q(\cdot \mid r_{<t})) \,,$$

Here, P and Q are the next-token distributions under the intervention and reference conditions, respectively, with their visual and question contexts omitted from the notation. Because the global view must be identified to the model, the additional-view condition includes a role instruction as part of the context intervention. This paired-prefix design measures the distributional change induced by each intervention. Figure 4(a) pools the resulting per-response scores across the sampled examples.

Spatial-reasoning rollout diversity. We compare View-AID and ViewFusion on the same sampled MMSI-Bench questions. For each question, we draw multiple responses from both models using identical decoding settings, with temperature 1.0 and top-p 0.9. We extract each model's spatial reasoning block and encode valid, non-empty traces with the same Qwen3 Embedding model (Zhang et al., 2025b). For a question q with $K_q \geq 2$ valid traces, let $e_1, \ldots, e_{K_q}$ denote their embeddings. Their average pairwise similarity is

$$S(q) = \frac{2}{K_q(K_q-1)} \sum_{i<j} \frac{e_i^\top e_j}{\|e_i\|_2 \|e_j\|_2}.$$

We average $S(q)$ equally across all paired questions in the sampled subset. This question-balanced aggregation measures reasoning-path diversity under a fixed semantic encoder, independently of agreement between final answers.