SPATIALASM: INTERNALIZING ACTIVE MENTALIZATION FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

When solving complex spatial reasoning tasks, ordinary humans naturally revisit the scenes from task-relevant viewpoints and mentally reconstruct the spatial layouts from those viewpoints. But can Large Vision-Language Models (LVLMs) proactively think in the same way? We term this ability **Active Spatial Mentalization (ASM)**, an intrinsic component of visual intelligence that remains substantially underdeveloped in current LVLMs. To bridge this gap, we introduce **SpatialASM**, a scalable framework to internalize ASM into LVLMs through training-time spatial simulation. SpatialASM employs an allocentric novel-view generator to ground hypothetical viewpoints in visual evidence and produces two complementary forms of supervision: **i)** explicit ASM trajectories that teach LVLMs when and how to adaptively invoke view planning and imagination during spatial reasoning, and **ii)** automatically-constructed perspective-taking tasks that further strengthen viewpoint-conditioned scene imagination. On seven benchmarks spanning relatively perception-oriented spatial understanding to more challenging tasks requiring multi-image, cross-view and viewpoint-transformation reasoning, SpatialASM yields performance gain on LVLMs of four LVLM backbones in spatial reasoning while retaining their original general vision capabilities. The experimental results demonstrate that our approach improves average accuracy across the seven spatial benchmarks. And the synthesis of ASM trajectories provides a scalable pathway to alleviate the scarcity of spatial Chain-of-Thought (CoT) supervision for eliciting active viewpoint exploration and mental scene simulation.

1 INTRODUCTION

Spatial reasoning is a foundational capability for Large Vision-Language Models (LVLMs), underpinning applications from autonomous driving and embodied navigation to robotic manipulation (Chen et al., 2024a; Cheng et al., 2024; Li et al., 2026c). While LVLMs achieve strong performance on many visual tasks (Xing et al., 2026; Liu et al., 2024b), they still fail on complex spatial problems that involve viewpoint transformation, cross-view correspondence, and reasoning about unobserved regions (Yang et al., 2026b; 2025; Fu et al., 2024; Xu et al., 2026). These tasks go beyond passively extracting relations from the given observations, as they require the model to reason about the three-dimensional scene from perspectives that the input images do not directly provide.

We argue that a core capability missing from current LVLMs is the ability to *proactively* determine which alternative viewpoints would reveal useful evidence and to construct what the scene looks like from those viewpoints mentally (Shepard & Metzler, 1971). We term this joint capability **Active Spatial Mentalization (ASM)**, comprising two complementary skills: **view planning**, which selects task-relevant viewpoints given the question and current observations, and **view imagination**, which mentally constructs the spatial layout expected from the selected viewpoints (Fig. 1(c)). Teaching ASM to LVLMs is difficult. Recent data-centric efforts have expanded spatial supervision through large-scale datasets built from 3D annotations and videos (Gao et al., 2026; Liu et al., 2026; Yang et al., 2026a), but these datasets provide answer-only question-answering (QA) pairs rather than Chain-of-Thought (CoT) trajectories that show *how* to select informative viewpoints and imagine the corresponding observations (Fig. 1(a)). Current LVLMs are unreliable at bootstrapping this supervision themselves, as they often produce incorrect viewpoint plans and hallucinated scenes when asked to generate ASM trajectories from scratch.

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.

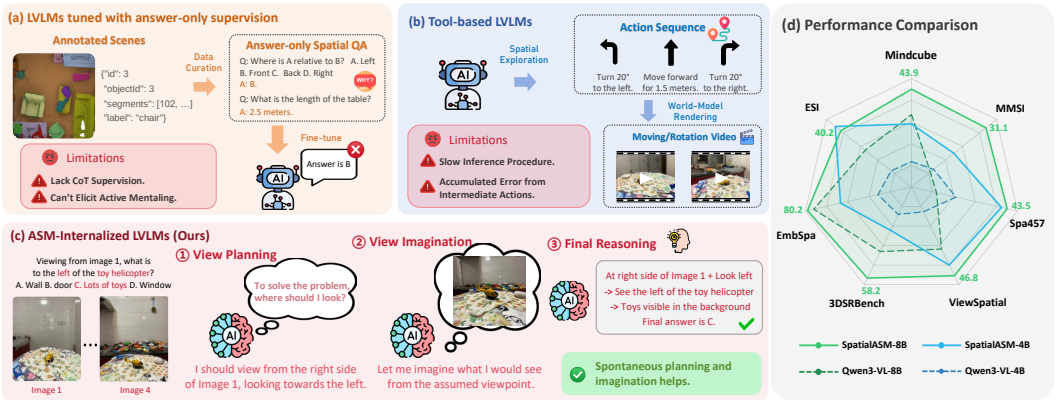


Figure 1: (a) Prior data-centric approaches mostly provide answer-only supervision. (b) Reasoning with external calls of world models adds unnecessary intermediate observations and extra inference time. (c) SpatialASM enables LVLMs to intrinsically produce plan-imagine-reason CoT. (d) Experiments demonstrate that SpatialASM delivers consistent performance gains on seven benchmarks.

An alternative is to attach an external world model or spatial tool at inference time to generate supplementary observations (Yang et al., 2026c; Cao et al., 2026; Zhang et al., 2026c) (Fig. 1(b)). While this provides useful visual evidence, it ties spatial reasoning to an external simulator at inference time, even when the LVLM learns when to invoke it (Yu et al., 2026; Zhu et al., 2026). Tool distillation (Zhou et al., 2026; Li et al., 2026d) removes this dependence, but targets perceptual tools such as depth and cognitive maps rather than viewpoint planning and view imagination. Most such methods also adopt an egocentric interface that reaches a target viewpoint through a sequence of camera-centric actions, so each query requires a multi-step rollout (Yang et al., 2026c; Cao et al., 2026). Closer to our setting, World2VLM (Zhang et al., 2026b) uses a world model at training time, but it supervises egocentric view association without producing any reasoning trajectory.

We introduce **SpatialASM**, a framework that equips LVLMs with ASM capabilities by using an allocentric novel-view generator (NVG) as a **training-time spatial simulation engine** rather than an inference-time tool (Fig. 1(c)). Given a spatial QA sample, the NVG renders views from question-conditioned viewpoints planned by the LVLM, grounding hypothetical perspectives in concrete visual evidence. SpatialASM transforms these into two complementary forms of supervision: explicit ASM reasoning trajectories for supervised fine-tuning (SFT) and automatically verifiable perspective-taking tasks for reinforcement learning (RL) with GRPO (Shao et al., 2024). The rendered views serve only as temporary scaffolding: in the final SFT data, they are replaced by text captions that describe the imagined observations, so the LVLM learns to construct spatial layouts internally from the original input images alone.

The design of SpatialASM is guided by two key mechanisms (Fig. 2). The **allocentric** viewpoint formulation specifies a target view by its location and direction in a single step, avoiding the multi-step rollouts of egocentric interfaces. A **Geometry Perturbation (GeoP)** metric provides a model-based proxy for the geometric consistency of rendered views; together with an answer-consistency check, it admits only samples that pass pre-defined thresholds into the training data. Across four backbones from the Qwen3-VL and Qwen3.5 series (Bai et al., 2025; Qwen Team, 2026a), SpatialASM improves average accuracy over seven spatial benchmarks (Tab. 1), with average gains ranging from $+1.2$ to $+3.5$ points. The largest per-benchmark gains appear on tasks that demand spatial mental modeling and viewpoint transformation, reaching $+8.7$ on Spatial457-6d and $+5.2$ on MindCube. General-VQA performance remains within 0.6 points of the baselines across all four backbones.

Our contributions are threefold. **1)** We formulate Active Spatial Mentalization as the joint capability of proactively planning task-relevant viewpoints and imagining corresponding observations, and argue that the absence of ASM is a key bottleneck of spatial reasoning in current LVLMs. **2)** We introduce SpatialASM, a scalable framework built on an allocentric novel-view generator and GeoP-based filtering that converts existing spatial QA data into explicit ASM trajectories for SFT and perspective-taking tasks for RL. **3)** SpatialASM improves average spatial accuracy on four LVLM backbones over seven benchmarks while keeping general-VQA performance within 0.6 points.

2 Related Work

We thoroughly discuss more related work here in Appendix A.

Proactive spatial intelligence. Current LVLMs still struggle with tasks requiring cross-view association, spatial mental modeling, and active evidence acquisition (Zhang et al., 2026a; Yang et al., 2026b). ESI-Bench (Hong et al., 2026) recasts spatial intelligence as a perception-action loop and shows that active exploration improves accuracy over passive observation. The two components of ASM have also been studied separately. For view imagination, APC (Lee et al., 2025) is a training-free pipeline that builds a 3D scene abstraction and re-projects it to a perspective specified in the question. For view planning, recent work trains LVLMs to choose camera viewpoints, either to localize a given target view (Wang et al., 2026) or to answer a question (Koo et al., 2025), but the chosen views are then rendered from a 3D environment rather than imagined. GeoThinker (Li et al., 2026b) internalizes a form of active reasoning at the feature-retrieval level, but such geometry-enhanced approaches (Zheng et al., 2026) require architecture-specific fusion modules that limit portability across LVLM architectures. In contrast, SpatialASM couples view planning and view imagination in a single question-conditioned reasoning trajectory that operates at the viewpoint and observation level and can be internalized through standard post-training without architectural modification.

Spatial data synthesis. Many studies enhance spatial reasoning by synthesizing large-scale datasets (Fig. 1(a)). Early approaches build on public scene datasets (*e.g.*, ScanNet (Dai et al., 2017)) with 3D annotations (Li et al., 2026c; Ouyang et al., 2025), or employ expert models (*e.g.*, segmentation, depth estimation) for data curation (Cheng et al., 2024; Deng et al., 2026). Recent data engines scale further: HoliSpatial (Gao et al., 2026) renders scenes through 3D Gaussian Splatting, and VST (Yang et al., 2026a) uses simulators to generate novel observations. Despite their effectiveness, the dominant supervision remains answer-only, which provides the final answer but little guidance on *how* to actively acquire or mentally construct missing evidence. SpatialASM addresses this gap by augmenting existing spatial QA pairs with ASM reasoning trajectories that couple question-conditioned view planning, viewpoint-conditioned imagination, and evidence-based reasoning.

World-model-assisted spatial reasoning. Drawing on generative world models (Team et al., 2025; Chen et al., 2026; Agarwal et al., 2026), several approaches equip LVLMs with external spatial simulators during test-time reasoning (Fig. 1(b)). MindJourney (Yang et al., 2026c) and SpatialDreamer (Cao et al., 2026) use video-generation models to synthesize observations under egocentric motions. Astra (Zhu et al., 2026) and AVIC (Yu et al., 2026) further train the LVLM to decide when and how much to imagine, but the imagined views are still produced by an external simulator at inference time. A second line uses world models or renderers only at training time. World2VLM (Zhang et al., 2026b) distills motion-conditioned view association from a world model; GeoWorld-VLM (Gu et al., 2026) aligns the vision encoder with world-model features under a frozen LLM; and Imaginative Perception Tokens (Bigverdi et al., 2026) train a unified model to generate latent views specified in the question, supervised by simulator renders. SpatialASM belongs to this second line. Its difference lies mainly in what is distilled: question-conditioned view planning and viewpoint-conditioned imagination, written out as explicit reasoning trajectories, so that LVLMs learn when and where to imagine rather than only how a view changes under a given motion. Test-time simulators and training-time internalization are complementary and could in principle be combined (Tab. 2).

3 SpatialASM

SpatialASM (Fig. 2) uses an allocentric novel-view generator (NVG) G as a training-time spatial simulator that converts answer-only spatial QA data into two complementary forms of supervision. First, G is fine-tuned to render the observation from a queried viewpoint, and we introduce a Geometry Perturbation (GeoP) metric to assess the geometric consistency of its outputs (Sec. 3.1). We then synthesize explicit ASM trajectories for SFT, in which a pretrained LVLM plans task-relevant viewpoints, G renders the corresponding views, and these views are then replaced by textual captions so that the model learns to imagine, rather than receive, the missing observations (Sec. 3.2). Finally, we construct automatically verifiable perspective-taking questions for reinforcement learning (RL), which expose the model to observations from diverse viewpoints and reinforce the association between a viewpoint description and its observation (Sec. 3.3). The trajectories supervise the full plan-imagine-reason procedure, while RL strengthens viewpoint-conditioned imagination.

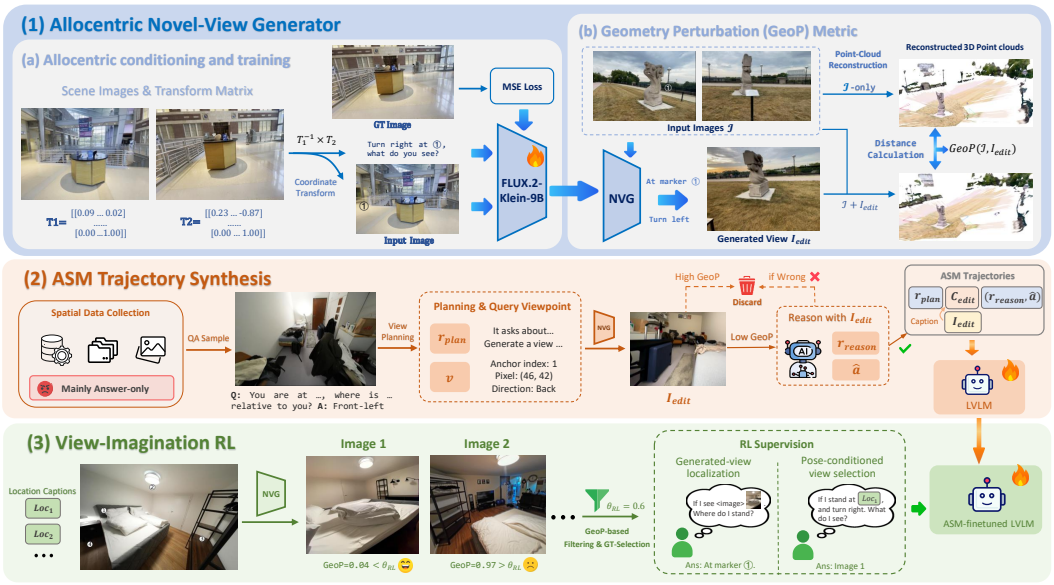


Figure 2: **Overview of SpatialASM. (1)** An allocentric novel-view generator (NVG, G) renders the scene from a queried viewpoint, and GeoP measures how much a rendered view perturbs the 3D reconstruction of the inputs. **(2)** An LVLM plans viewpoints, G renders them, and filtered trajectories are converted to captions for SFT. **(3)** GeoP-filtered candidate views form verifiable perspective-taking questions for RL.

3.1 Allocentric Novel-View Generator

Allocentric conditioning. We obtain SpatialASM-NVG (denoted as G) by fine-tuning a pretrained image-editing model to render the observation from a queried viewpoint (Fig. 2(1a)). Most existing world models act as egocentric simulators: a viewpoint change is expressed as a sequence of ego-motion commands with quantified translations and rotations. In contrast, G is conditioned directly on the target viewpoint, so that a task-relevant perspective can be queried in a single step.

Given a set of images $\mathcal{I} = \{I_1, \ldots, I_n\}$ of the same scene, we define an **allocentric viewpoint** as $v = (i, x, y, d)$, where (x, y) is a pixel coordinate on the anchor image I_i that specifies the target observation location and d is the viewing direction. The anchor index i is required because a pixel coordinate is only meaningful relative to one image, and providing the full set $\mathcal{I}$ gives G additional scene context beyond I_i. Given $\mathcal{I}$ and v, G renders the observation from the queried viewpoint:

$$I_{\text{edit}} = G(\mathcal{I}, v), \quad v = (i, x, y, d). \tag{1}$$

Training. Each training sample consists of the image set $\mathcal{I}$, an allocentric viewpoint $v = (i, x, y, d)$, and the ground-truth observation I_{gt} from that viewpoint. We fine-tune the image-editing model on these samples, supervising the rendered view $I_{\text{edit}} = G(\mathcal{I}, v)$ with the ground-truth observation I_{gt}. This supervision asks G to preserve the scene content of $\mathcal{I}$ while changing the perspective as specified by v. Details of the data construction are given in the Appendix C.1.

Geometry Perturbation (GeoP) metric. Standard novel-view synthesis metrics such as PSNR, SSIM, and LPIPS measure image-level similarity to a reference view and therefore do not capture whether a rendered view is geometrically consistent with the scene under large viewpoint changes.

To assess the geometric plausibility of views rendered by G, we introduce GeoP, a model-based metric that measures how much a rendered view perturbs the 3D geometry inferred from the original observations (Fig. 2(1b)). Given the image set $\mathcal{I}$, we first reconstruct the scene with a pretrained 3D reconstruction model F, obtaining a point cloud $\mathcal{M}_{\text{orig}} = F(\mathcal{I})$ from the original observations alone. We then add the rendered view $I_{\text{edit}} = G(\mathcal{I}, v)$ to the observations and run the reconstruction:

$$\mathcal{I}^+ = \mathcal{I} \cup \{I_{\text{edit}}\}, \quad \mathcal{M}_{\text{aug}} = F(\mathcal{I}^+). \tag{2}$$

We define $\mathrm{GeoP}(\mathcal{I}, I_{\mathrm{edit}})$ as the Chamfer distance between $\mathcal{M}_{\mathrm{orig}}$ and $\mathcal{M}_{\mathrm{aug}}$:

$$\mathrm{GeoP}(\mathcal{I}, I_{\mathrm{edit}}) = \frac{1}{|\mathcal{M}_{\mathrm{orig}}|} \sum_{\mathbf{p}\in\mathcal{M}_{\mathrm{orig}}} \min_{\mathbf{q}\in\mathcal{M}_{\mathrm{aug}}} \|\mathbf{p}-\mathbf{q}\|_2^2 + \frac{1}{|\mathcal{M}_{\mathrm{aug}}|} \sum_{\mathbf{q}\in\mathcal{M}_{\mathrm{aug}}} \min_{\mathbf{p}\in\mathcal{M}_{\mathrm{orig}}} \|\mathbf{q}-\mathbf{p}\|_2^2. \tag{3}$$

I_{edit} with a lower GeoP is better in quality as it is more geometrically consistent with the input image set $\mathcal{I}$. We use GeoP to filter rendered views with thresholds (Sec. 3.2 and Sec. 3.3) and to compare generators in our experiments. We explain more details of the GeoP metric in Appendix B.

3.2 ASM Trajectory Synthesis

Motivation. Existing spatial datasets offer large-scale supervision, but most contain only question-answer (QA) pairs without intermediate reasoning. Such answer-only supervision does not teach a model how to perform ASM on spatial tasks that require reasoning beyond the given observations. Meanwhile, current LVLMs often select uninformative viewpoints and imagine novel views unreliably (Sec. 4.1). Directly prompting an LVLM to write ASM trajectories therefore risks inaccurate or hallucinated intermediate steps, which would yield misleading supervision.

Trajectory synthesis. We therefore synthesize ASM trajectories automatically, using G as an external spatial simulator during data construction only (Fig. 2(2)). From large-scale open-source spatial datasets, we select subtasks that require active perspective exploration beyond direct recognition, *e.g.*, reasoning about camera-object relationships and spatial configurations, associating multiple input views, and inferring what is seen from a viewpoint other than the original ones.

Each QA sample from the selected subtasks is a triplet $(q, \mathcal{I}, a)$, where q is the question, $\mathcal{I}$ is the set of input images, and a is the ground-truth answer. For each sample, we prompt a pretrained LVLM M to produce a view plan with two parts: a reasoning trace r_{plan} that infers which viewpoints would help answer q, and K allocentric viewpoints $\{v_k\}_{k=1}^K$ that can be used as the condition of G:

$$(r_{\mathrm{plan}}, \{v_k\}_{k=1}^K) = M(q, \mathcal{I}), \quad v_k = (i_k, x_k, y_k, d_k). \tag{4}$$

The planned viewpoints $\{v_k\}$ specify where and in which direction the model should observe the scene to answer q. G then renders a view $I_{\mathrm{edit},k} = G(\mathcal{I}, v_k)$ for each of them.

The rendered views are returned to M together with a textual description t_{des} of the planned viewpoints, and M answers the question again, producing a reasoning trace r_{reason} and an answer $\hat{a}$:

$$(r_{\mathrm{reason}}, \hat{a}) = M\big(q, t_{\mathrm{des}}, \mathcal{I}, \{I_{\mathrm{edit},k}\}_{k=1}^K\big). \tag{5}$$

Filtering. We retain a trajectory only if the rendered views are both useful and geometrically consistent: M must answer incorrectly from $\mathcal{I}$ alone but correctly after seeing the rendered views, and every rendered view must have a GeoP below a threshold θ. Let $\hat{a}_0$ denote the answer of M from $\mathcal{I}$ alone. It is checked first, so the set of retained image-form trajectories for a question $(q, \mathcal{I})$ is

$$\mathcal{T}_{\mathrm{img}} = \big\{(r_{\mathrm{plan}}, \{I_{\mathrm{edit},k}\}_{k=1}^K, r_{\mathrm{reason}}, \hat{a}) \,\big|\, \hat{a}_0 \neq a,\ \hat{a} = a,\ \max_k \mathrm{GeoP}(\mathcal{I}, I_{\mathrm{edit},k}) < \theta\big\}, \tag{6}$$

Caption substitution. To internalize view imagination rather than view consumption, we replace each rendered view in a retained trajectory with a textual caption. A captioning model M_{cap} describes $I_{\mathrm{edit},k}$ as $C_{\mathrm{edit},k} = M_{\mathrm{cap}}(I_{\mathrm{edit},k})$ and is prompted to focus on the 2D layout and relative positions of objects in the view. The final ASM trajectories for a question $(q, \mathcal{I})$ are

$$\mathcal{T}_{\mathrm{ASM}} = \big\{(r_{\mathrm{plan}}, \{C_{\mathrm{edit},k}\}_{k=1}^K, r_{\mathrm{reason}}, \hat{a}) \,\big|\, (r_{\mathrm{plan}}, \{I_{\mathrm{edit},k}\}_{k=1}^K, r_{\mathrm{reason}}, \hat{a}) \in \mathcal{T}_{\mathrm{img}}\big\}. \tag{7}$$

Each trajectory thus supervises the model to plan informative viewpoints for the task, to imagine the observation and spatial layout from each planned viewpoint in text, and to reason with the imagined views. Because the rendered images never appear in the supervision, the model learns to carry out this procedure from the original observations alone and does not call G at inference time. This supervision targets multi-step spatial tasks that require information beyond the original observations.

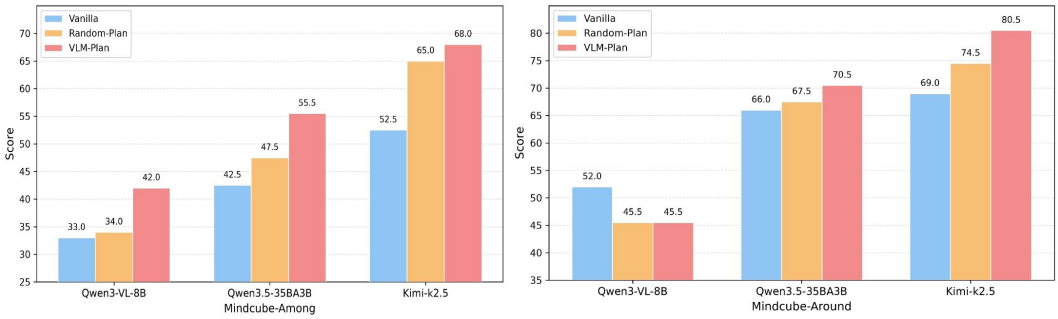


Figure 3: **Evaluation of external ASM**. We evaluate vanilla LVLM reasoning, and two external ASM strategies with different viewpoint-planning methods. Both planning-based settings invoke SpatialASM-NVG and condition the final reasoning on the original and generated observations.

3.3 View-Imagination Reinforcement Learning

Motivation. To further strengthen the ability of LVLMs to imagine a scene from a designated allocentric viewpoint, we construct automatically verifiable multiple-choice questions and train the SFT model from Sec. 3.2 on them with GRPO (Shao et al., 2024) (Fig. 2(3)). These questions train the model to associate an allocentric viewpoint query with its observation, without manually annotated reasoning traces.

Candidate views. For an image set $\mathcal{I}$ of a scene, we randomly sample four allocentric viewpoints $\{v_k = (1, x_k, y_k, d_k)\}_{k=1}^{4}$ anchored on I_1 and render one candidate view for each with G:

$$\mathcal{V} = \{I_{\text{edit},1}, \ldots, I_{\text{edit},4}\}, \quad I_{\text{edit},k} = G(\mathcal{I}, v_k). \tag{8}$$

From each candidate set $\mathcal{V}$ we build two complementary multiple-choice tasks. *1) Generated-view localization*. We overlay four numbered markers on the anchor image I_1 at the sampled locations $\{(x_k, y_k)\}$ to obtain I_{mark}, and pick one candidate $I_{\text{edit},j}$. Given $(I_{\text{mark}}, \mathcal{I}, I_{\text{edit},j})$, the model must determine at which numbered marker the candidate view was captured and the answer is j. *2) Pose-conditioned view selection*. This task is the inverse: given a textual target pose, the model must pick the matching candidate. For a chosen index j (selected as described below), we use M_{cap} to describe the location (x_j, y_j) on I_1 and combine this description with the viewing direction d_j into a target-pose specification t_j. Given t_j, $\mathcal{I}$, and the four candidates $\mathcal{V}$, the model must select the candidate that depicts the scene from the described location and direction and the answer is j.

Answer selection. The ground-truth candidate $I_{\text{edit},j}$ must be geometrically consistent with $\mathcal{I}$ for the reward to be meaningful, so we select j for each candidate set $\mathcal{V}$ with GeoP:

$$\mathcal{K} = \{k \mid \text{GeoP}(\mathcal{I}, I_{\text{edit},k}) < \theta_{\text{RL}}\}, \quad j = \begin{cases} \text{any } k \in \mathcal{K}, & \mathcal{K} \neq \emptyset, \\ \text{discard the sample}, & \mathcal{K} = \emptyset, \end{cases} \tag{9}$$

where θ_{RL} is a pre-defined threshold. The two tasks provide complementary supervision: *generated-view localization* learns the inverse mapping from an observation to its allocentric viewpoint, whereas *pose-conditioned view selection* learns the forward mapping from a textual target pose to its expected observation. Both answers are derived automatically without human annotation, and the two tasks are intended to improve the ability of LVLMs to imagine and reason about scene layout and object configuration from viewpoints other than the observed ones.

4 Experiments

Dataset construction. To train the NVG in Sec. 3.1, we leverage *DL3DV* (Ling et al., 2024) to construct our training set **SpatialASM-edit140k**, containing both single-image and multi-image inputs. To synthesize ASM trajectories introduced in Sec. 3.2, we source various challenging spatial-QA samples from existing spatial datasets and generate a dataset with 128k reasoning trajectories, denoted **SpatialASM-128k**, and in practice, we use the subset **SpatialASM-66k** constructed exclusively from *Holispatial* for tuning LVLMs. For RL (Sec. 3.3), we automatically construct **SpatialASM-RL-58k** by employing the NVG. More details are provided in Appendix C.

Table 1: **Performance comparison** on 7 spatial reasoning and 3 general-VQA benchmarks. We report the average accuracy over all benchmarks within each category. The performance of several spatial models are also provided (7B variants on Qwen2.5-VL-7B; 8B variants on Qwen3-VL-8B).

Models	Spatial Reasoning								Gnr.-VQA
	Mindcube	MMSI	Spa457_{6d}	Viewspa	3DSR	EmbSpa	ESI	Avg.	Avg.
Representative Spatial Models									
SpaceR-7B	36.6	28.8	45.1	38.4	54.5	37.3	40.2	40.1	59.4
VST-RL-7B	39.3	35.1	43.4	47.0	60.1	72.0	40.1	48.1	62.1
HoliSpatial-8B	52.0	30.6	30.5	50.1	49.7	73.8	41.8	46.9	63.1
SenseNova-8B	50.4	29.6	41.3	46.4	55.2	70.8	36.5	47.2	63.7
Baselines & Our Model (Qwen3-VL-4B)									
Qwen3-VL-4B	33.9	28.0	37.2	43.2	55.8	72.3	36.5	43.8	63.7
SpatialASM-4B	39.1	29.2	42.8	46.2	56.4	76.5	40.6	47.3	63.1
Improvement	+5.2	+1.2	+5.6	+3.0	+0.6	+4.2	+4.1	**+3.5**	-0.6
Baselines & Our Model (Qwen3-VL-8B)									
Qwen3-VL-8B	40.4	27.8	34.8	45.3	57.2	79.5	38.4	46.2	64.4
SpatialASM-8B	43.9	31.1	43.5	46.8	58.2	80.2	40.2	49.1	65.0
Improvement	+3.5	+3.3	+8.7	+1.5	+1.0	+0.7	+1.8	**+2.9**	**+0.6**
Baselines & Our Model (Qwen3-VL-30BA3B)									
Qwen3-VL-30BA3B	36.7	25.2	41.9	43.9	54.5	73.4	37.2	44.7	69.4
SpatialASM-30BA3B	37.9	30.8	42.1	45.6	58.8	75.2	40.3	47.2	68.9
Improvement	+1.2	+5.6	+0.2	+1.7	+4.3	+1.8	+3.1	**+2.5**	-0.5
Baselines & Our Model (Qwen3.5-9B)									
Qwen3.5-9B	50.8	33.5	41.4	47.3	64.4	81.4	39.1	51.1	78.1
SpatialASM-9B	54.7	35.2	43.5	47.9	63.5	82.3	39.2	52.3	78.5
Improvement	+3.9	+1.7	+2.1	+0.6	-0.9	+0.9	+0.1	**+1.2**	**+0.4**

Training setup. To internalize ASM capabilities into LVLMs, we adopt the two-stage post-training strategy: first SFT then GRPO. During SFT, we train four models from both the Qwen3-VL and the Qwen3.5 series (Bai et al., 2025; Qwen Team, 2026a) on SpatialASM-66k for 10 epochs with the learning rate of 5e-6. For GRPO, we use a global batch size of 128 and a learning rate of 1e-6, and train for 300 steps. Training details are provided in Appendix D.1.

Evaluation setup. We evaluate all models on seven Spatial Reasoning (SR) benchmarks and three general-VQA (Gnr.-VQA) benchmarks. All benchmarks except ESI-Bench are evaluated using VLMEvalKit (Duan et al., 2024). For ESI-Bench, we adopt the passive-single setting. We ask the models to produce reasoning content first on the SR tasks, and to answer directly on the general-VQA tasks. Benchmark and evaluation details are provided in Appendix D.2 and Appendix D.3.

4.1 Preliminary Analysis of Explicit ASM

Before internalizing ASM into LVLMs, we first wonder whether ASM is an effective reasoning paradigm and whether existing LVLMs lack this capability. Like previous studies in Fig. 1(b), we conduct a preliminary experiment on the external ASM settings, where LVLMs first plan question-relevant viewpoints, invoke the NVG to synthesize observations from these viewpoints, and then reason jointly over the original and generated images. We sample 200 questions from each of the *Among* and *Around* subtasks of MindCube and report the results of three reasoning methods: (i) Vanilla: LVLMs directly reason over the original observations; (ii) Random-Plan: randomly selected target viewpoints are passed to SpatialASM-NVG, and (iii) VLM-Plan: LVLMs are explicitly prompted to plan question-relevant viewpoints before invoking SpatialASM-NVG.

By comparing *Vanilla* with the other two settings in Fig. 3, the planning-based external ASM paradigm generally improves accuracy over vanilla reasoning, which demonstrates that viewpoint planning and imagination constitute a generally effective paradigm for complex spatial reasoning. Moreover, LVLM-based planning consistently matches or outperforms random planning, indicating that existing LVLMs already possess a certain ability to identify informative viewpoints when

Table 2: **Accuracy and average inference time of different methods on benchmark subtasks.** (i) Vanilla: the LVLM reasons directly from the input observations alone, and the reasoning duration is reported; (ii) + External ASM: Similar to Sec. 4.1, the LVLM externally invokes SpatialASM-NVG to help reasoning. The time to synthesize novel-view observations is reported.

Subtask	Qwen3-VL-8B				SpatialASM-8B			
	Vanilla		+ External ASM		Vanilla		+ External ASM	
	Accuracy	T(s)	Accuracy	T_{NVG}(s)	Accuracy	T(s)	Accuracy	T_{NVG}(s)
$\text{Mindcube}_{\text{Among}}$	33.0	12.6	42.0	57.6	39.5	18.6	**45.0**	56.8
$\text{Mindcube}_{\text{Around}}$	52.0	7.5	45.5	38.1	58.0	7.2	**61.5**	46.8

explicitly prompted. Nevertheless, this ability is not spontaneously activated during direct LVLM reasoning (*Vanilla* setting). In other words, existing LVLMs are able to execute view-planning only under an explicitly-prompted inference pipeline. ASM has not yet become an intrinsic reasoning capability that they can autonomously invoke when needed for certain complicated spatial tasks.

Furthermore, Qwen3-VL-8B drops under both random and VLM-based planning, suggesting that relatively small models still lag behind in both robust planning for diverse spatial tasks and long-context multimodal reasoning. It further motivates internalizing ASM through training so that models can learn when and how to employ this reasoning paradigm more reliably.

4.2 Main Results

Overall SR performance gains. As shown in Table 1, SpatialASM improves spatial reasoning on both Qwen3-VL and Qwen3.5 architectures. The Qwen3-VL models of three different scales, including dense models and an mix-of-experts (MoE) model, yield consistent performance gains on all seven SR benchmarks after applying our approach. In particular, the 4B model achieves an average improvement of **+3.5%**, with notable gains on Mindcube (+5.2%) and Spatial457 (+5.6%). In contrast, prior spatial models exhibit uneven performance across benchmarks. For example, Sensenova-8B improves significantly on Mindcube (+10.0%), but exhibits drops by 8.7% on EmbSpa. This suggests that their gains may depend more on benchmark-specific task formats, whereas SpatialASM elicits a more universal and transferable spatial capability.

Analysis of spatial capabilities. The largest gains generally occur on more challenging benchmarks (e.g., MindCube, MMSI, and $\text{Spatial457}_{\text{6d}}$) which require the comprehensive use of multiple reasoning capabilities, such as spatial mental modeling, cross-view correspondence, complex 6D reasoning, etc. These tasks benefit from the learned ability to plan informative viewpoints and imagine the corresponding observations. SpatialASM also yields modest gains on relatively perception-centric SR benchmarks (e.g., 3DSR, Embspa), indicating that the knowledge learned from ASM reasoning trajectories and the RL tasks improves the general understanding of the 3D space.

General Visual Capabilities. SpatialASM preserves the general visual capabilities of LVLMs, with changes in the average general-VQA accuracy remaining within 0.6% across all four backbones. Importantly, SpatialASM-8B outperforms the two 8B counterparts in *representative spatial models*, where HoliSpatial experiences an average degredation of 1.3% on all three benchmarks after SFT. We attribute this result partly to view-imagination RL, which exposes LVLMs to diverse visual observations and high-quality, geometry-filtered NVG outputs. Consequently, the model learns from rich visual supervision rather than merely memorizing the reasoning patterns of the SFT trajectories. The models' performance on each benchmark is shown in Appendix D.4.

Comparison with external ASM. As shown in Table 2, external ASM is not always beneficial for the baseline model. Introducing redundant visual evidence or unnecessary view-planning CoT may degrade performance (52.0% to 45.5% on *Around*). In contrast, after internalization, SpatialASM-8B improves under vanilla reasoning on both subtasks without noticeable increase on the inference duration. **Most importantly**, combining SpatialASM-8B with external ASM yields the highest accuracy on both Among (45.0%) and Around (61.5%). SpatialASM strengthens the model's intrinsic ability to plan viewpoints and reason about spatial evidence, while external NVG provides additional informative views. Employing them together achieves higher accuracy than either approach alone, suggesting that internalization and external tool-use provide complementary gains.

Table 3: Comparison of different novel-view generators under external ASM paradigm.

Methods	Acc_{among} ↑	Acc_{around} ↑	GeoP↓
Qwen3.5-35BA3B	42.5	66.0	/
+GPT-Image2	42.0	69.0	0.141
+SpatialEdit	46.0	68.0	0.110
+SVC	48.5	68.5	**0.087**
+SpatialASM-NVG	**55.5**	**70.5**	0.118
Kimi-k2.5	52.5	69.0	/
+GPT-Image2	52.0	71.0	0.157
+SpatialEdit	52.5	69.0	0.138
+SVC	52.5	74.0	**0.072**
+SpatialASM-NVG	**68.0**	**80.5**	0.099

Table 4: Ablations on training strategies and GeoP-based data filtering strategies.

Methods	Mindcube	EmbSpa	ViewSpa	ESI	Avg.
Qwen3-VL-4B	33.9	72.3	43.2	36.5	46.5
SFT with ASM trajectories					
w/o filter	36.4	75.0	45.3	36.8	48.4
with filter	38.7	74.6	**46.2**	40.2	49.9
View-imagination RL					
w/o filter	34.2	75.7	43.8	39.2	48.2
with filter	35.1	76.3	43.6	40.1	48.8
SFT + RL					
SpatialASM-4B	**39.1**	**76.5**	**46.2**	**40.6**	**50.6**

4.3 Ablation Studies

Comparison of novel-view generators. Following the experimental design in Sec. 4.1, we execute the external ASM reasoning paradigm with different novel-view generators and evaluate on the same Mindcube subsets as in Sec. 4.1. The accuracy as well as the median GeoP value of all generated images are reported. As shown in Table 3, compared with other world models, SpatialASM-NVG achieves **the highest accuracy** for both backbone LVLMs on both Mindcube subsets.

Interestingly, SpatialASM-NVG does not achieve the best GeoP as SVC obtains lower GeoP values for both LVLMs, but it consistently yields substantially higher reasoning accuracy. This observation suggests that **geometric consistency alone** does not fully determine the utility of generated views for downstream spatial reasoning. We attribute this to our allocentric form of viewpoint representation. Rather than describing the target viewpoint through a sequence of quantified egocentric motions, LVLMs can designate the viewpoint through a single query, and the corresponding location of the viewpoint is explicitly marked in the original input image. It makes the spatial correspondence between the original scene and the generated observation easier for LVLMs to understand during final multi-image reasoning. Therefore, our allocentric-view generation provides not only reliable novel views, but also auxiliary evidence that is more directly aligned with the reasoning objective.

Training and Data recipe. Table 4 compares the independent effect of each training stage and evaluates the filtering strategy for data preparation. All of the four single-stage variants are initialized from the baseline LVLM. For both SFT and RL, applying GeoP-based filtering improves the average performance on the four SR benchmarks, indicating that GeoP effectively removes geometrically distorted novel views and prevents unreliable visual evidence from being incorporated into ASM reasoning trajectories or RL supervision.

The two training stages also exhibit **distinct but complementary** effects. SFT with SpatialASM-66k (with filter) produces strong gains on tasks requiring **explicit viewpoint transformation** and multi-step spatial reasoning, such as MindCube (+4.8%) and ViewSpa (+3.0%), which verifies the high quality of our ASM trajectories and the reliability of the proposed CoT synthesis engine from answer-only spatial QAs. In contrast, view-imagination RL is especially effective on the more **perception-oriented** EmbSpa benchmark and outperforms the setting of applying SFT alone. Combining SFT and RL achieves the best overall accuracy, suggesting a clear synergy where SFT primarily teaches the structured ASM reasoning pattern while RL further strengthens viewpoint-conditioned spatial perception and imagination. Some detailed design is given in Appendix D.5.

5 Conclusion

In this work, we identify Active Spatial Mentalization (ASM) as a key bottleneck in LVLM spatial reasoning. To address this issue, we propose SpatialASM, a novel post-training framework that internalizes ASM capabilities into LVLMs through training-time spatial simulation. Experiments demonstrate that SpatialASM consistently improves spatial reasoning while preserving general visual capabilities. More importantly, our framework provides a scalable approach to constructing CoT supervision for existing spatial datasets, addressing the common lack of intermediate reasoning processes in open-source spatial data.

AI USE STATEMENT

In this work, we used generative AI tools for minor language polishing only. We have not used generative AI tools for research ideation, data synthesis, or mathematical claims proving, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all words and expressions polished by AI in the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is completely built upon publicly available datasets and models. To the best of our knowledge, our entire implementation does not raise concerns regarding discrimination, bias, fairness, privacy, security, or legal compliance. No new data involving human subjects are collected for this study.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we will publicly release our dataset, model checkpoints, and code (including training scripts and data synthesis framework) upon publication.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Mahtab Bigverdi, Linjie Li, Weikai Huang, Yiming Liu, Jaemin Cho, Tuhin Kundu, Chris Dongjoo Kim, Zelun Luo, Jieyu Zhang, Linda Shapiro, et al. Imaginative perception tokens enhance spatial reasoning in multimodal language models. In *European Conference on Computer Vision*, pp. 40–57. Springer, 2026.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025.

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

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7176–7187, 2026.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Xinyuan Chen, Haoyu Guo, Shi Guo, Bingqi Jiang, Chunhua Shen, Xing Shen, Tianfan Xue, Yufei Xue, Mulin Yu, Weinan Zhang, et al. A definition and roadmap for world models. *arXiv preprint arXiv:2607.06401*, 2026.

Zhixun Chen, Ming Li, Yuxuan Huang, Yali Du, Meng Fang, and Tianyi Zhou. Atlas: Agent tuning via learning critical steps. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25334–25349, 2025.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, et al. Spatialclaw: Rethinking action interface for agentic spatial reasoning. *arXiv preprint arXiv:2606.13673*, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Baoliang Tian, Dingwen Zhang, et al. S-agent: Spatial tool-use elicits reasoning for spatial intelligence. *arXiv preprint arXiv:2606.20515*, 2026.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Jinhui Yin, Qi Wei, Tianshuo Yang, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 50773–50814, 2026.

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

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 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. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, et al. Holi-spatial: Evolving video streams into holistic 3d spatial intelligence. *arXiv preprint arXiv:2603.07660*, 2026.

Renjie Gu, Kaichen Zhou, Yan Luo, and Mengyu Wang. GeoWorld-VLM: Geometry from world models for vision-language models. *arXiv preprint arXiv:2605.16713*, 2026.

Zefeng He, Xiaoye Qu, Yafu Li, Tong Zhu, Siyuan Huang, and Yu Cheng. Diffthinker: Towards generative multimodal reasoning with diffusion models. *arXiv preprint arXiv:2512.24165*, 2025.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Li Fei-Fei, Jiajun Wu, and Yejin Choi. Esi-bench: Towards embodied spatial intelligence that closes the perception-action loop. *arXiv preprint arXiv:2605.18746*, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

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.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Juil Koo, Daehyeon Choi, Sangwoo Youn, Phillip Y Lee, and Minhyuk Sung. Toward ambulatory vision: Learning visually-grounded active view selection. *arXiv preprint arXiv:2512.13250*, 2025.

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

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision*, pp. 95–111. Springer, 2026a.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, JiaWang Bian, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026b.

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. In *International Conference on Learning Representations*, volume 2026, pp. 76566–76592, 2026c.

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*, 2026d.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Jianhui Liu, Haoze Sun, Wenbo Li, Yanbing Zhang, Rui Yang, Zhiliang Zhu, Yijun Yang, Shenghe Zheng, Nan Jiang, Jiaxiu Jiang, et al. Openspatial: A principled data engine for empowering spatial intelligence. *arXiv preprint arXiv:2604.07296*, 2026.

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

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

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3DSRBench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025.

OpenAI. GPT Image 2 model, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Juneyoung Ro, Namwoo Kim, and Yoonjin Yoon. How well do vision-language models understand cities? a comparative study on spatial reasoning from street-view images. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6535–6544. IEEE, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 1971.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Kangrui Wang, Linjie Li, Zhengyuan Yang, Shiqi Chen, Zihan Wang, Li Fei-Fei, Jiajun Wu, Leonidas Guibas, Lijuan Wang, and Manling Li. Planning with the views. *arXiv preprint arXiv:2605.29563*, 2026.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

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

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24669–24679. IEEE, 2025c.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, 2024.

Yicheng Xiao, Wenhu Zhang, Lin Song, Yukang Chen, Wenbo Li, Nan Jiang, Tianhe Ren, Haokun Lin, Wei Huang, Haoyang Huang, et al. Spatialedit: Benchmarking fine-grained image spatial editing. *arXiv preprint arXiv:2604.04911*, 2026.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 13066–13093, 2026.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. Citycube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215, 2026.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. In *European Conference on Computer Vision*, pp. 192–211. Springer, 2026a.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026b.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *Advances in Neural Information Processing Systems*, 38:109855–109885, 2026c.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *arXiv preprint arXiv:2602.08236*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026a.

Wanyue Zhang, Wenxiang Wu, Wang Xu, Jiaxin Luo, Helu Zhi, Yibin Huang, Shuo Ren, Zitao Liu, and Jiajun Zhang. World2vlm: Distilling world model imagination into vlms for dynamic spatial reasoning. *arXiv preprint arXiv:2604.26934*, 2026b.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1724–1732. IEEE, 2024.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, et al. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026c.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025.

Yang Zhou, Zixuan Huang, Sunzhu Li, Zhuo Yang, Chen Zhang, Shunian Chen, Caijun Yan, Jianyao Xu, Shunyu Liu, Weijie Fu, et al. Spatialcli: Learning to reason with spatial tools, then without them. *arXiv preprint arXiv:2607.27703*, 2026.

Chenming Zhu, Jingli Lin, Yilin Long, Peizhou Cao, Tai Wang, Jiangmiao Pang, and Xihui Liu. Thinking with imagination: Agentic visual spatial reasoning with world simulators. *arXiv preprint arXiv:2606.06476*, 2026.

A More Related Work

A.1 Proactive Spatial Intelligence

With the progress of LVLMs, recent benchmarks are moving beyond static spatial recognition (Kamath et al., 2023; Wang et al., 2025b) and understanding towards more demanding spatial reasoning tasks where LVLMs still perform significantly worse than ordinary human beings. For example, Spatial457 (Wang et al., 2025c) progressively evaluates spatial understanding from object recognition and 2D localization to complex 6D reasoning, revealing a pronounced performance degradation as geometric complexity increases. MindCube (Wang et al., 2025a) studies whether LVLMs can construct a coherent spatial mental model from only a few observations, which evaluates cognitive mapping, perspective taking, and mental simulation of hypothetical movements. Similarly, MMSI-Bench (Yang et al., 2026b) targets multi-view spatial intelligence and identifies scene reconstruction, cross-view correspondence, and perspective-transformation reasoning as major failure modes of current LVLMs. These benchmarks suggest that successful spatial reasoning often requires more than extracting relations from the observations already provided: the model must reason about how the scene would be perceived under alternative unobserved viewpoints.

The recent ESI-Bench (Hong et al., 2026) delves deeper into the **perception-action loop** and treat it as a higher level of spatial intelligence. An agent can selectively employ locomotion, and manipulation to uncover information that cannot be resolved from the initial observation alone. By reasoning under the *active multiple* setting, where the models can proactively execute movements and interact with the simulation environments, the performance significantly outperforms that under the *passive single* setting. However, the accuracy of the *active multiple* setting lags far behind the *GT passive* setting, where the best ground-truth passive trajectories are provided. These results suggest that current LVLMs lack intrinsic proactive exploration capability when solving spatial problems, unless they are explicitly prompted to do so. And their trajectory-planning capability is still far from saturated.

To internalize the thinking pattern of 'perception-active loop' into LVLMs, Geothinker (Li et al., 2026b) proposes active geometry integration, in which semantic representations selectively query task-relevant geometric features. However, such geometry-enhanced approaches (Zheng et al., 2026; Li et al., 2026b) introduce an additional pretrained feature branch and architecture-specific fusion modules into the LVLMs, and require alignment between geometric and semantic representations, potentially limiting their adaptability across heterogeneous VLM architectures. Unlike their notion of 'active' at the **feature-retrieval level**, our ASM operates at the **viewpoint and observation level**. It reasons about *where* useful evidence should be acquired and *what* should be observed from the corresponding hypothetical viewpoint. And we focus on data-centric pipeline where such capability can be internalized through off-the-shelf post-training strategies without introducing geometry module or making architectural modification.

A.2 SPATIAL DATA SYNTHESIS

A wide spectrum of previous work lay emphasis on large-scale high-quality data curation pipeline to improve spatial reasoning in LVLMs. A straightforward approach is to construct such supervision via human annotation (Cai et al., 2025), but it is hardly affordable to scale. Some other pipelines resort to proprietary/open-source large-language models (Ro et al., 2025; Cheng et al., 2024). However, current models' performance on relatively challenging spatial reasoning tasks, especially in multi-view or video modalities, remains far from satisfactory and may introduce abundant errors to curated QA pairs.

To formulate scalable spatial supervision with high accuracy and quality, various synthesis frameworks emerge with two predominant characteristics: leveraging public scene datasets with explicit geometric annotations, and extensively employing visual expert models to recover otherwise unavailable 3D information from ordinary images. **i)** Numerous approaches (e.g., SpatialRGPT, OpenSpatial) leverage open-source scene datasets (*e.g.*, ScanNet (Dai et al., 2017), ARKitScenes (Baruch et al., 2021)) with diverse annotated scenes as their foundational data (Li et al., 2026c; Ouyang et al., 2025; Liu et al., 2026). These datasets provide 3D annotations, such as object classification, semantic voxel labels, and 3D segmentations, which can be easily augmented to spatial QA tasks. **ii)** A complementary direction further adds necessary annotations for constructing spatial supervision through specialized visual models (Chen et al., 2024a; Cheng et al., 2024; Deng et al., 2026). These models have grown relatively mature in the traditional field of computer vision and perform well in tasks such as segmentation (Liu et al., 2024a), metric depth estimation (Hu et al., 2024), object detection (Zhang et al., 2024), and pose estimation (Wang et al., 2024), thereby lifting 2D imagery to 3D context.

While these pipelines greatly expand the scale of spatial supervision, the resulting data are largely constrained by the expert models or the information already encoded in the source public datasets. Recent works therefore explore richer sources of supervision. For instance, VST (Yang et al., 2026a) broadens the coverage of spatial supervision across single-image, multi-image, and video scenarios by leveraging a 3D scene simulator to generate novel observations. Holi-Spatial (Gao et al., 2026) turns raw video streams into geometrically optimized 3D Gaussian Splatting scenes (Kerbl et al., 2023) and automatically derives multi-level supervision (e.g., 3D grounding instances, QA pairs). SenseNova-SI (Cai et al., 2026) scales spatial intelligence to a new level, curating millions of spatial training samples by incorporating and filtering QA pairs from an incredibly large variety of datasets.

Despite the increasing scale and diversity of these datasets, most of them only provide **answer-only** supervisory signals. Specifically, for those relatively challenging tasks (e.g., perspective-taking allocentric transformation, mental reconstruction), the answers remain centered on **what** spatial relation or camera motion should be predicted, rather than explicitly demonstrating **how** a model can approach the correct answer through proactively acquiring missing evidence during reasoning.

VST-R (Yang et al., 2026a) constitutes an important step towards reasoning supervision by providing 135K curated sample with explicit step-by-step reasoning traces. However, its synthesis process requires ground-truth 3D bounding boxes in the original scene data to form intermediate Bird's-Eye View (BEV) images. Moreover, its reasoning supervision primarily teaches the model to derive spatial conclusions from the provided visual evidence, rather than to execute the *perception-action loop* of proactive exploration. It is not yet publicly released as well. In contrast, SpatialASM grounds an explicit question-conditioned reasoning process of *when* additional viewpoint reasoning is needed, *where* the model should mentally observe, and *what* evidence should be expected from that viewpoint, which is more demanding and human-like. Importantly, SpatialASM can synthesize ASM reasoning trajectories for existing spatial QA data without requiring any prerequisite ground-truth 3D annotations.

World2VLM (Zhang et al., 2026b) is particularly relevant to our work in that it also employs a generative world model in its data engine and distills its spatial imagination into the LVLMs. World2VLM synthesizes future views based on given parameterized camera trajectories to teach LVLMs to associate multi-view observations with egocentric motion transitions. Though effective in enhancing view-imagination capabilities, World2VLM fails to teach how to initialize a thinking process with imaginary world-model-like motions when encountering spatial problems, nor does it provide any reasoning content in its supervision. Our objective is different: SpatialASM starts from the **spatial question itself** and teaches the model to determine which viewpoint would provide useful missing

evidence, rather than starting from a predefined egocentric motion trajectory and learning the corresponding visual outcome. The generated views serve only as temporary scaffolding for synthesizing **SpatialASM-128k**, whose final supervision contains long-CoT traces which explicitly couples **question-conditioned view planning**, **viewpoint-conditioned imagination**, and **reasoning** with the imagined evidence.

A.3 Tool-Assisted Spatial Reasoning

Recent studies increasingly augment LVLMs with paradigms of external harness to improve spatial intelligence. The tools can be primarily categorized into three groups: specialized visual tools, 3D reconstruction tools, and generative world models.

Specialized visual tools are computer-vision models which provide more reliable low-level or metric spatial evidence, such as pose/depth estimation and segmentation. Multiple studies demonstrate that the accurate specialist perception substantially improves spatial reasoning when LVLMs are specifically optimized for determining which tools to invoke and how to integrate their outputs (Cho et al., 2026; Dai et al., 2026; Zhou et al., 2026). Pretrained **3D reconstruction** models recover point clouds and camera poses from scene images (Keetha et al., 2026; Lin et al., 2025) so that LVLMs can actively manipulate the reconstructed scene and inspect spatial configurations from alternative viewpoints (Zhang et al., 2026c). A third direction employs **generative world models** to synthesize observations that are not present in the original visual inputs (Zhu et al., 2026; Yu et al., 2026; Yang et al., 2026c). These methods primarily enhance spatial reasoning by augmenting the external evidence, which introduce additional inference overhead due to external tool execution and LVLM-tool interaction. Moreover, specialized tool-use training on LVLMs may introduce an additional trade-off between tool proficiency and the original general capabilities (Chen et al., 2025).

SpatialASM targets a fundamentally different yet complementary dimension from the above tool-assisted approaches: strengthening the intrinsic spatial reasoning process itself. While tool-assisted methods improve the availability and precision of external spatial evidence, SpatialASM improves the model's ability to reason about what spatial evidence is needed and how it should be used. The two paradigms can thus be **naturally combined**, improving spatial intelligence along complementary axes of intrinsic reasoning capability and external evidence accessibility.

B Geometric Perturbation (GeoP) Metric

Geometric Perturbation (GeoP) is designed to assess the geometric plausibility of views generated by the NVGs, especially under a relatively large perspective transformation. The calculating method is provided in Sec. 3.1. Some details are further illustrated in this section. Though $\mathcal{M}_{\text{orig}}$ and $\mathcal{M}_{\text{aug}}$ are obtained through two separate feed-forward inference process of F, both point clouds are aligned in the same world coordinate system from its predicted depth maps, camera intrinsics, and camera poses, so they can be compared directly. In terms of **the implication of GeoP**, a low GeoP means that adding I_{edit} changes the reconstructed geometry little, *i.e.*, F can explain the rendered view as another observation of the same scene. A high GeoP means that the rendered view forces F to alter its scene-level geometry, which we take as a sign of geometric inconsistency. GeoP is therefore a proxy that depends on F rather than a ground-truth measure.

C Dataset Details

C.1 SpatialASM-edit140k

C.1.1 Statistics

SpatialASM-edit140k consists of 139,730 samples, all automatically constructed from *DL3DV* (Ling et al., 2024), a public dataset which provides 10k diverse scenes captured from multiple viewpoints together with the corresponding camera matrices. As illustrated in Sec. 3.1, each sample has the form of $(\mathcal{I}, v, I_{\text{gt}})$, containing the input images $\mathcal{I}$, the allocentric viewpoint v, and expected ground-truth novel-view observation. To enable G to generate novel views from multi-image inputs, we set the length of input images $|\mathcal{I}|$ to different values, and the number of samples of each $|\mathcal{I}|$ is provided in Tab. 5.

Table 5: The samples numbers of each input-image length.

Input Length	$\lvert\mathcal{I}\rvert = 1$	$\lvert\mathcal{I}\rvert = 2$	$\lvert\mathcal{I}\rvert = 3$	$\lvert\mathcal{I}\rvert = 4$	**Total**
Size	35,026	34,835	34,846	35,023	139,730

C.1.2 Construction Method

Each training sample has a full image set $\mathcal{I}$, an allocentric query viewpoint $v = (i, x, y, d)$, and the ground-truth observation I_{gt} from that viewpoint. In practice, (x, y) is provided to G as a visual marker drawn on I_i, and d as a textual prompt. We build such samples from public scene datasets that provide multiple posed observations of the same scene. For a pair of frames from one scene, we take one frame as the anchor I_i and the other as I_{gt}; the relative camera pose between them determines where the camera of I_{gt} lies in I_i and which direction it faces, which yields (x, y, d). The detailed construction process is introduced as follows.

Given a 3D scene video in $DL3DV$, we first select a few observations as $\mathcal{I} = \{I_1, ..., I_n\}$, and another frame as I_{gt}. Critically, I_{gt} is excluded from $\mathcal{I}$ and the frame indices of any two images in $\mathcal{I}$ differ by at least 30, ensuring a noticeable viewpoint difference between them.

Since the image-editing model cannot directly accept the viewpoint query $v = (i, x, y, d)$ as its input, we separate and integrate it into the image and prompt inputs. In practice, we lay a visual marker on the pixel (x, y) of the anchor image I_i, and we transform d into a textual prompt. Here we will explain how we derive (x, y) from the metadata provided in $DL3DV$.

Assume that I_1 is the anchor image we select, then we need to compute the viewpoint of I_{gt} to derive (x, y) in the query based on the coordinate system of I_1. We can easily collect their corresponding camera-to-world transformation matrices $T_1, T_{\text{gt}} \in SE(3)$ from the original dataset. $SE(3)$ denotes the set of 3D rigid transformations, each represented by a 4×4 homogeneous matrix that combines a 3D rotation and a translation. The transformation from camera-gt to camera 1 is:

$$T_{gt\to 1} = T_1^{-1} T_{gt} = \begin{bmatrix} R_{gt\to 1} & \mathbf{t}_{gt\to 1} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix}, \tag{10}$$

where $R_{gt\to 1} \in \mathbb{R}^{3\times 3}$ is the relative rotation of camera-gt with respect to camera 1, and $\mathbf{t}_{gt\to 1} = [t_x, t_y, t_z]^T$ is the position of camera-gt in the coordinate system of camera 1. This position is projected onto I_1 as:

$$x = f_x \frac{t_x}{t_z} + c_x, \qquad y = f_y \frac{t_y}{t_z} + c_y, \tag{11}$$

where (f_x, f_y) are the focal lengths and (c_x, c_y) is the principal point, also provided in $DL3DV$ originally. A valid projection requires $t_z > 0$ and the resulting pixel (x, y) to lie inside the image boundary. Finally, we derive the allocentric viewpoint direction d based on the relative rotation around the y-axis, which can be computed from $R_{gt\to 1}$ and expressed as a pitch angle in degrees. And we assign d to the corresponding direction in $\{\text{forward}, \text{left}, \text{back}, \text{right}\}$.

C.1.3 Training samples

We provide two examples with single and multiple image inputs in Fig. 4. As in some training samples, when the queried viewpoint reveals regions outside the views of the original visual inputs, we allow and encourage G to imagine objects **not present** in the input images, provided that these imagined objects are plausible in the real world (e.g., the patio umbrella and the window on the left wall in I_{gt} of Case 2).

C.2 SpatialASM-128k

In Sec. 3.2, we explain how to synthesize the ASM reasoning trajectories based on triplets $(q, \mathcal{I}, a)$ from existing answer-only spatial datasets. We construct 128k ASM trajectory samples, and use a subset of 66k samples for our SFT training. Our approach provides a scalable and convenient pathway to synthesizing high-quality spatial CoTs as it does not require any other metadata (e.g., BEV image, bounding box). It helps address the scarcity of reasoning content in prior large-scale spatial supervision and elicit active task-conditioned planning and imagination capabilities in LVLMs from scratch.



Figure 4: Examples in SpatialASM-edit140k. (Case 1, 2: $|\mathcal{I}| = 1$, Case 3: $|\mathcal{I}| = 2$)

C.2.1 Statistics

All original spatial-QAs are sourced from Holi-Spatial (Gao et al., 2026), VST-500K (Yang et al., 2026a), and the training set of MindCube (Wang et al., 2025a). We give the number of samples from each original data source in SpatialASM-128k in Tab. 6. Importantly, to avoid distribution overlap between training data and our final evaluation on benchmarks, SpatialASM-66k **does not contain** any sample synthesized from MindCube-train.

Table 6: The number of samples from each data source in SpatialASM-128k

Data Source	Holi-Spatial	VST-500k	MindCube-train	**Total**
Sample No.	91,648	34,741	1,965	128,354

C.2.2 Synthesis Protocols

Some detailed implementations for ASM trajectory synthesis are given in this section. In Sec. 3.2, we have mentioned that we use two pretrained LVLMs, M and M_{cap}, to generate certain parts in our CoT trajectories. M is prompted to produce planning trace r_{plan} and the final reasoning trace r_{reason}, and M_{cap} is used to transform $\{I_{\text{edit},k}\}_{k=1}^{K}$ into $\{C_{\text{edit},k}\}_{k=1}^{K}$. In practice, we employ Qwen3.6-35BA3B (Qwen Team, 2026c) as M and Qwen3.6-27B (Qwen Team, 2026b) as M_{cap}. For filtering, θ is used to discard samples whose $\{I_{\text{edit},k}\}_{k=1}^{K}$ have high GeoP values. Since we also filter for reasoning correctness in SFT (Equation 6), we adopt a relatively lenient quality threshold for $\{I_{\text{edit},k}\}_{k=1}^{K}$. We set $\theta = 2.5$ during our filtering process.

Here we provide the system prompts of **planning**, **reasoning**, and **captioning**. The planning prompt is for producing $(r_{\text{plan}}\{v_k\}_{k=1}^{K})$ in Equation 4; the reasoning prompt is for producing $(r_{reason}, \hat{a})$ in Equation 5; the captioning prompt is for producing $\{C_{\text{edit},k}\}_{k=1}^{K}$ in Equation 7.

- **Planning prompt**: I gave you several images and a spatial reasoning question. You can now use an image-editing model to generate images of the scene from new perspectives based on the input images.

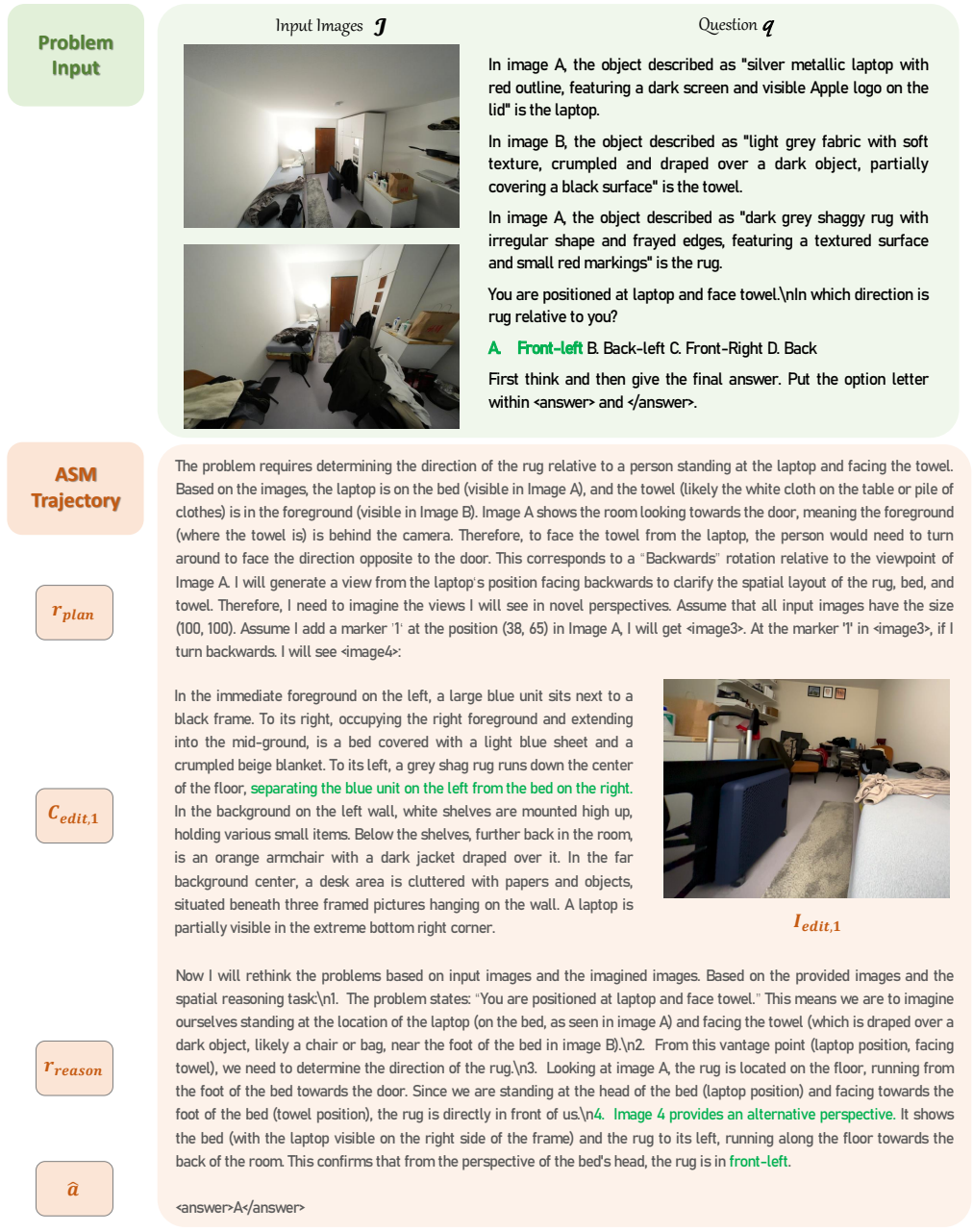


Figure 5: Example in SpatialASM-128k. Note that $I_{\text{edit, 1}}$ does not exist in the ASM trajectory.

You do not need to answer the spatial reasoning question I gave you, but you need to reason based on the question and give commands to the image-editing model to provide more useful information for answering this question.

For each command, you need to first select a source image from the input images. Then you need to select an area on the image and assign a facing direction, so that the image-editing model will generate the scene viewed from this area and specific facing direction.

Provide your answer in the following format: a list containing Dict(s):

[{'source_img': '<e.g. 1>', 'pixel_x': '<e.g. 22>', 'pixel_y': '<e.g. 63>', 'Direction': 'e.g. <Right>'}, ...]

'source_img' is the anchor image number (the first image is numbered 1, and so on). 'pixel_x' and 'pixel_y' represent the assigned area, and we assume the width and height

of the images are both 100 for simplicity. Direction is chosen from 'Left', 'Backwards', 'Forward' and 'Right'.

You can output an empty list if you believe there is no need for providing a new perspective.

First think and then give the final answer. The final answer should contain a short summary of your reasoning process (no more than 200 words), which only focus on how to derive the command without giving the answer of the original question, and the final list.

The summary should be put within <summary> and </summary>, and then output the list within <answer> and </answer>.

- **Reasoning prompt**: First think and then give the final answer. Put the option letter within <answer> and </answer>.
- **Captioning prompt**: Describe the 3D location of the objects in the image. Locations include left, right, top, bottom, background, foreground, and so on. You can describe the location of each object separately and the relative location between objects as well (e.g. Object A is to the right of Object B). The answer should be no more than 200 words.

C.2.3 DATA SAMPLE

We present a sample of SpatialASM-128k in Fig. 5. The sample contains the original problem input $(\mathcal{I}, q)$ and the ASM trajectory $(r_{\text{plan}}, \{C_{\text{edit},k}\}_{k=1}^{K}, r_{\text{reason}}, \hat{a})$. In this case, the number of auxiliary imagined view is $K = 1$. Critically, the generated novel views only act as scaffolding for conversion into $\{C_{\text{edit},k}\}_{k=1}^{K}$, and do not really exist in the final dataset.

C.3 SPATIALASM-RL-58K

We construct the SpatialASM-RL-58k automatically by employing SpatialASM-NVG without any LVLM/human annotation. Only the original 2D images come from Holi-Spatial (Gao et al., 2026) and DL3DV (Ling et al., 2024). SpatialASM-RL-58k consists of 58,346 training samples, with 28,594 from Holi-Spatial and 29,752 from DL3DV.

In Sec. 3.3 we introduce the *answer selection* strategy to reduce potential ambiguity in the two multi-choice tasks and ensure the quality of the rendered images. A pre-defined parameter θ_{RL} is applied in this strategy. During the construction process, we find that generated images generally have higher GeoP values when the queried allocentric viewpoint is set to *backwards* than when it is set to other directions. This is because the backward direction is more likely to reveal regions not visible in the input images, requiring the NVG to synthesize more unobserved content. Therefore, even when the generated images exhibit similar levels of geometric consistency, those generated for the backward direction tend to receive higher GeoP. After manually inspecting a large number of generated images and their corresponding GeoP values, we set θ_{RL} to 1.2 for images whose corresponding target direction is *backwards*, and 0.6 for all other directions.

In Fig. 6, we give example VQAs of the two tasks we construct for RL: *generated-view localization* and *pose-conditioned view selection*. The formulations of both tasks are illustrated in Sec. 3.3.

D EXPERIMENT DETAILS

D.1 TRAINING DETAILS

SpatialASM-NVG training. We finetune FLUX.2-klein-9B (Black Forest Labs, 2025), an image-editing model for real-time text-conditioned image generation, on SpatialASM-edit140k to derive our SpatialASM-NVG. Drawing on the training strategy in DiffThinker (He et al., 2025), we use mean squared error (MSE) as our training objective and train for 3 epochs. We only finetune its Diffusion Transformer (DiT) with Low-Rank Adaptation (LoRA) and we set LoRA rank to 512.

LVLM Training. We fully fine-tune Qwen3-VL-4B and Qwen3-VL-8B, while applying LoRA (rank=32) to Qwen3-VL-30B-A3B and Qwen3.5-9B. During GRPO, we apply KL regularization with weight 1e-2. And we sample five responses in each group with the temperature set to 1.0. Our reward is composed of two items: accuracy reward and format reward, with the weight of 0.9 and 0.1. Accuracy reward is a binary reward which is assigned to 1 only when the final option is correct.

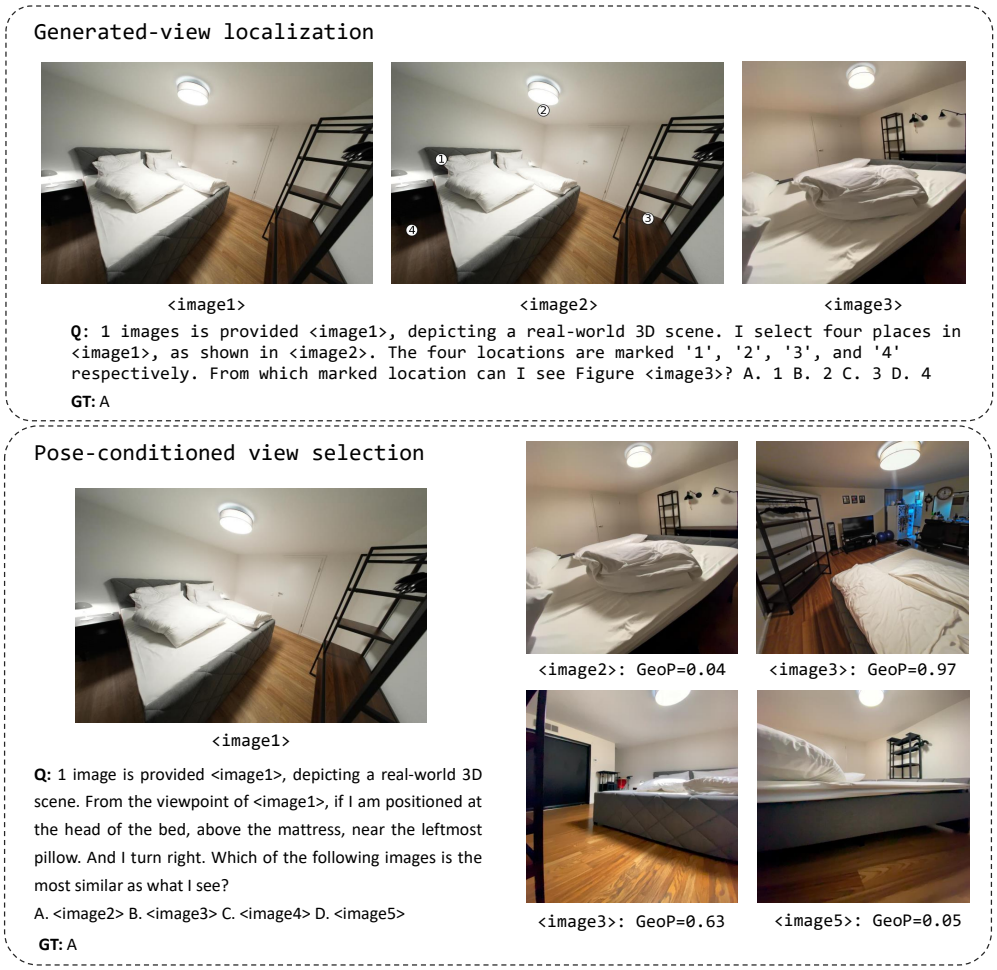


Figure 6: Examples of view-imagination RL tasks.

Format reward is also binary which requires the LVLMs to reason first and put the option letter within <answer> and </answer>.

Prompt. For both SFT and RL, we append a system prompt to the end of the original question, guiding LVLMs to output first reasoning content and then the final answer. The system prompt is: *First think and then give the final answer. Put the option letter within <answer> and </answer>.*

D.2 Benchmark Overview

Spatial Reasoning. We evaluate all models on seven Spatial Reasoning (SR) benchmarks (Wang et al., 2025a; Yang et al., 2026b; Wang et al., 2025c; Li et al., 2026a; Ma et al., 2025; Du et al., 2024; Hong et al., 2026). For Spatial457 (Wang et al., 2025c), we only select its most challenging 6d subtask with 1837 questions. **ESI-Bench** (Hong et al., 2026), **MindCube** (Wang et al., 2025a), **MMSI-Bench** (Yang et al., 2026b), and **Spatial457-6d** are the most difficult among the seven benchmarks, requiring LVLMs to go beyond directly reading spatial relations from the given observations and instead integrate incomplete evidence, mentally construct a 3D scene representation, and reason about unobserved or hypothetical viewpoints. **ViewSpatialBench** (Li et al., 2026a) represents an intermediate level of difficulty, with camera-perspective questions which can often be answered directly from the observed coordinate frame and person-perspective questions which require some basic perspective-taking capabilities. Finally, **3DSRBench** (Ma et al., 2025) and **EmbSpatial-Bench** (Du et al., 2024) are relatively perception-oriented, covering spatial properties such as height, orientation, location, and egocentric spatial relations. Taken together, these benchmarks provide a compre-

Table 7: Performance comparison on three general-VQA benchmarks.

Models	MMMU_{VAL}	MMStar	RealWorldQA	Average
Representative Spatial Models				
SpaceR-7B	51.8	61.4	65.0	59.4
VST-RL-7B	51.2	63.7	71.4	62.1
HoliSpatial-8B	56.0	65.4	68.0	63.1
SenseNova-8B	55.1	64.5	71.6	63.7
Baselines & Our Model (Qwen3-VL-4B)				
Qwen3-VL-4B	54.7	65.1	71.3	63.7
SpatialASM-4B	53.8	64.0	71.6	63.1
Baselines & Our Model (Qwen3-VL-8B)				
Qwen3-VL-8B	54.1	66.0	73.2	64.4
SpatialASM-8B	56.1	67.1	71.8	65.0
Baselines & Our Model (Qwen3-VL-30BA3B)				
Qwen3-VL-30BA3B	62.7	70.7	74.8	69.4
SpatialASM-30BA3B	62.1	68.6	76.1	68.9
Baselines & Our Model (Qwen3.5-9B)				
Qwen3.5-9B	73.3	78.6	82.4	78.1
SpatialASM-9B	75.3	78.5	81.7	78.5

hensive evaluation of spatial reasoning across complementary dimensions, ranging from fundamental spatial perception and egocentric relation recognition to perspective transformation, multi-view scene reconstruction, and imagination-driven reasoning about unobserved spaces.

General VQA. We additionally evaluate on **three** general-VQA benchmarks (Yue et al., 2024; Chen et al., 2024b; xAI, 2024) to examine whether SpatialASM preserves the broad visual capabilities of the base LVLMs. $\textbf{MMMU}_{\textbf{val}}$ (Yue et al., 2024) evaluates expert-level multimodal understanding and reasoning across 6 major disciplines. **MMStar** (Chen et al., 2024b) evaluates 18 detailed ability axes in 6 major categories, spanning fine-grained and coarse visual perception, instance-level reasoning, and science and technology, with an emphasis on genuine visual dependency. **RealWorldQA** (xAI, 2024) focuses on basic visual understanding in real-world scenes, including counting, object size comparison, relative direction, optical character recognition (OCR), etc. Together, these benchmarks measure whether the spatial improvements introduced by SpatialASM come at the expense of general visual understanding.

D.3 Evaluation Details

Evaluation prompt. In Tab. 1, we evaluate the LVLMs on two major task categories: spatial reasoning and general-VQA (Gnr.-VQA). For spatial reasoning, we apply the same system prompt as training (in Appendix D.1) to let LVLMs reason before giving the answer. For general-VQA benchmarks, we apply another system prompt to enable direct responses of final answers from LVLMs: *Directly output your answer without any reasoning. The correct option letter is:*.

Representative spatial models. Four popular spatial LVLMs of previous studies are also evaluated on the benchmarks. SpaceR-7B (Ouyang et al., 2025) and VST-RL-7B (Yang et al., 2026a) are trained upon Qwen2.5-VL-7B (Team, 2025). Their official repository provides system prompts. For the SR evaluation, we follow them exactly to guide the models to produce structured reasoning and final answers. Holi-Spatial (Gao et al., 2026) and SenseNova-SI-1.1-Qwen3-VL-8B (Cai et al., 2026) are trained upon Qwen3-VL-8B Bai et al. (2025), similar to our SpatialASM-8B. Since the two models do not provide official system prompt, we follow our system prompt during evaluation.

D.4 General VQA Results

In Tab. 1, we only report the average score on the three general-VQA benchmarks. We provide the detailed accuracy of all LVLMs on each benchmark in Tab. 7.

D.5 Ablation Details

In Fig. 3 and Tab. 3, we employ the reasoning paradigm of external ASM: LVLMs execute question-conditioned view-planning, then invoke the NVG to synthesize corresponding observations, and finally reason based on the previous auxiliary information. Three different LVLMs are evaluated in Fig. 3: Qwen3-VL-8B (Bai et al., 2025), Qwen3.5-35BA3B (Qwen Team, 2026c), and Kimi-k2.5 (Team et al., 2026). And four different rendering models are compared in Tab. 3: GPT-Image-2 (OpenAI, 2026), SpatialEdit (Xiao et al., 2026), SVC (Zhou et al., 2025), and our SpatialASM-NVG.

In Tab. 3, GeoP is used for comparing the synthesis quality of different NVGs. For each setting, we report the median GeoP of the images generated by the NVG rather than the average. This is because most generated images have low GeoP values (below 1), while a small number of poor-quality images have much higher values (above 10). These outliers can disproportionately raise the average GeoP, making it less representative of the geometric consistency of a typical novel-view observation generated by a certain NVG/World-Model.

Tab. 4 ablates training and data-filtering strategies described in Sec. 3.2 and Sec. 3.3. For SFT, *w/o filter* means omitting the trajectory **filtering** step in Sec. 3.2 and using all synthesized ASM trajectories for training. For RL, *w/o filter* means omitting the GeoP-based criterion in the **answer selection** step of Sec. 3.3: no sample is discarded because its candidate views exceed the GeoP threshold, and the ground-truth candidates are selected from the full candidate set. Importantly, the final **SFT+RL** setting in Tab. 4 as well as all our models shown in Tab. 1 experience both training stages and all training data experience the filtering strategy.

E More Cases

E.1 SpatialASM-NVG Rendering

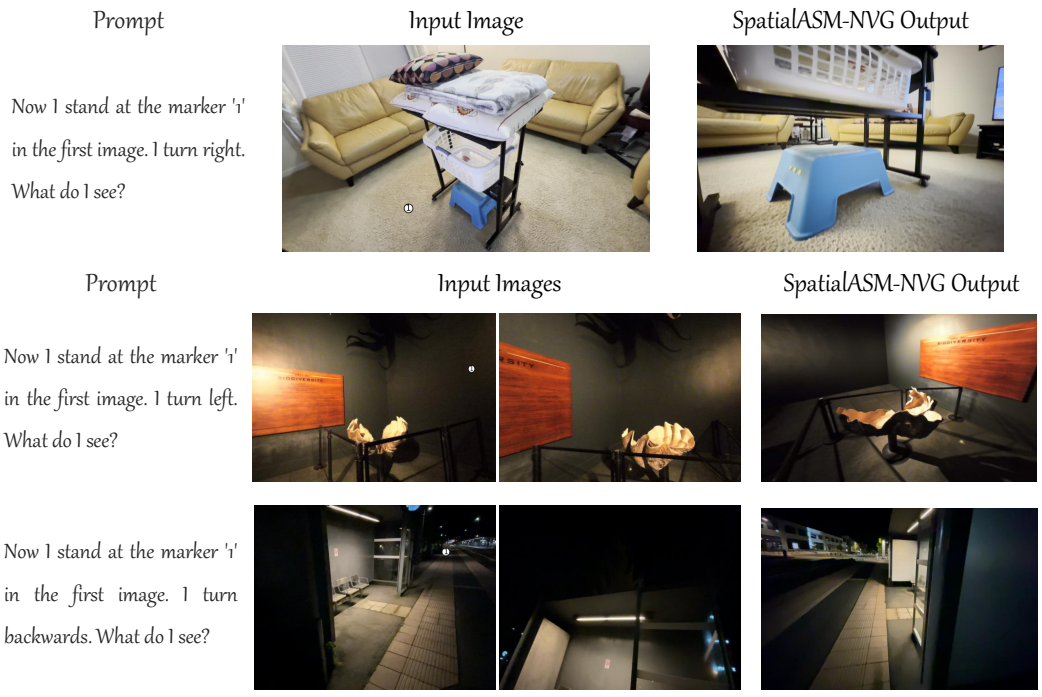


Figure 7: Examples of novel-view observations rendered by SpatialASM-NVG.

We obtain SpatialASM-NVG after training FLUX.2-klein-9B (Black Forest Labs, 2025) on SpatialASM-edit140k. Here we provide three output examples rendered by SpatialASM-NVG in Fig. 7.

Case 1: Turning right at the target location reveals an area that was originally outside the input image's field of view (beyond the sofa on the far right). SpatialASM-NVG successfully reconstructs **the missing content** in this region while retaining the objects in the original input (e.g., blue chair, white box).

Case 2: The generated view is not only geometrically consistent with the input but also faithfully reproduces the **lighting** on the brown panel on the right and the **shadows** cast on the floor.

Case 3: This case is indeed rather tricky, as the relative camera poses of the two input images are not immediately apparent. Nevertheless, our model correctly infers their spatial relationship and **seamlessly integrates** the tactile paving from Image 1 (invisible in Image 2) with the white panel on the inner left side of the bus stop from Image 2 (not visible in Image 1) into a single rendered image.

E.2 Results of Novel-view Generators



Figure 8: Case 1: Observations generated by different novel-view generators.

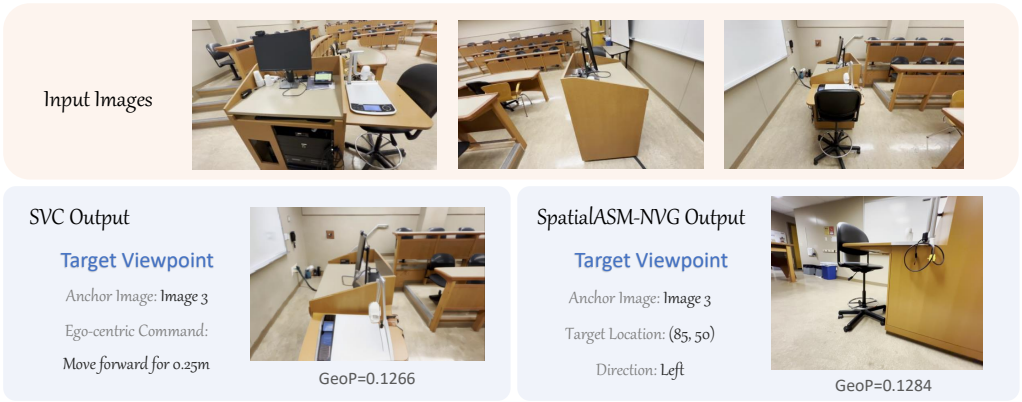


Figure 9: Case 2: Observations generated by different novel-view generators.

To showcase qualitative examples of the rendering results of different novel-view generators, we select some generated observations from the experiment in Tab. 3 with Kimi-k2.5 as the LVLM for planning and reasoning. The generated images are shown in Fig. 8 and Fig. 9. Evidently, the formulation of target viewpoints varies greatly. Prior approaches(e.g., SVC (Zhou et al., 2025)) use ego-centric motions from the camera position of an anchor image, whereas SpatialASM-NVG defines the target location directly by marking a pixel on an image. Note that in both figures we assume that the input images all have the size $(100, 100)$.

Case 1 (Fig. 8): The query target viewpoints of the two novel-view generators are approximately the same. SpatialASM-NVG has a much lower GeoP value because the red fire-fighting unit and

the black disinfectant unit in the output image more closely resemble their counterparts in the input images.

Case 2 (Fig. 9): Although SVC's output image is more geometrically consistent with the visual inputs (lower GeoP), its egocentric formulation results in relatively small camera motions and its output remains highly similar to the anchor image, providing little additional information for further LVLM reasoning. In contrast, our allocentric formulation induces larger viewpoint changes and provides more informative output observations, which could also possibly help the LVLM better understand the relative camera poses across input images and the spatial layout of the 3D scene.

E.3 Examples of LVLM Internalization

Person Perspective

Q: From the perspective of the man, where is the dog positioned?
A. Front B. back-right C. left D. back



SpatialASM-8B:

To determine the dog's position relative to the man, we need to consider the perspective of the man sitting in the boat.
- The man is seated at the helm (steering position) of the small boat.
- The dog is standing on the deck of the boat, near the front (bow) of the vessel.
- From the man's seated position, the front of the boat is directly in front of him.
- The dog is positioned ahead of him, which corresponds to the "front" of the boat from his viewpoint.
Therefore, from the man's perspective, the dog is positioned in front of him.
<answer>A</answer>

Qwen3-VL-8B:

From the perspective of the man in the boat, the dog is standing on the right side of the boat, which would be the back-right from his viewpoint as he is facing forward. The man is seated at the helm, and the dog is standing near the stern (back) of the boat, to his right side.
<answer>B. back-right</answer>

Figure 10: Qualitative comparison of Qwen3-VL-8B and SpatialASM-8B.

We have trained several LVLMs to internalize ASM capabilities and improved their overall spatial reasoning performance (Tab. 1). Here, we provide **four qualitative comparisons** between SpatialASM-8B and its baseline, Qwen3-VL-8B. These examples show that SpatialASM-8B can adaptively decide whether to employ the ASM reasoning pattern in its output according to the task type and difficulty, rather than simply overfitting to the ASM reasoning trajectories. In fact, we find that the model does not explicitly follow the ASM pattern in most cases, yet its overall spatial reasoning accuracy still improves substantially (Tab 1). This suggests that our training **internalizes generalizable spatial reasoning capabilities** instead of imposing a rigid reasoning template.

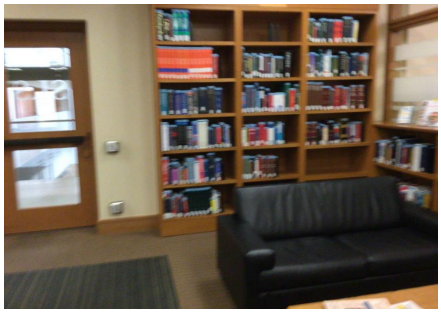
Fig. 10 and Fig. 11 show two problems targeting the viewpoint of *person* and *camera* respectively. Though SpatialASM-8B does not produce the ASM reasoning pattern explicitly, it accurately interprets the 3D spatial configuration from different perspectives and correctly answers both questions. Fig. 12 and Fig. 13 are more challenging tasks where SpatialASM-8B generates CoT of active spatial planning and imagination to help reach the correct answer.

Camera Perspective

Q: Where is the sofa in relation to the door?
A. back-up B. above-right C. front-down D. back-right

Qwen3-VL-8B:

The door is on the left side of the image. The sofa is on the right side of the image, positioned in front of the bookshelves. Therefore, relative to the door, the sofa is to the right and in front. The option that best describes this position is "front-down" (C), as it implies the sofa is in front of the door and slightly lower in the frame (which is consistent with the perspective).
<answer>C</answer>

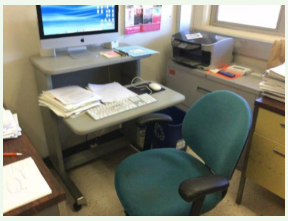


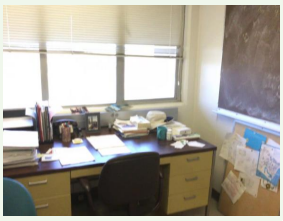
SpatialASM-8B:

To determine the position of the sofa relative to the door, we need to analyze their spatial relationship in the image.
1. **Identify the Door:** The door is located on the left side of the image.
2. **Identify the Sofa:** The black leather sofa is located on the right side of the image, occupying the lower-right portion of the frame.
3. **Determine Relative Position:** **From the perspective of the viewer looking at the image**, the sofa is to the right of the door. It is also positioned further into the room than the door, which is closer to the foreground. Therefore, the sofa is located "to the right and back" relative to the door.
Looking at the options:
……
- C. front-down: The sofa is not in front of the door from the viewer's perspective; it's to the side and back.
- D. back-right: This correctly describes the sofa's position relative to the door – it is behind (back) and to the right of the door.
<answer>D</answer>

Figure 11: Qualitative comparison of Qwen3-VL-8B and SpatialASM-8B.

Q: If the direction a person is facing while sitting on the black chair **is considered the front**, in which direction is the computer relative to the black chair? A. Behind B. In front C. Right D. Left





Qwen3-VL-8B:

To determine the direction of the computer relative to the black chair, we need to analyze the spatial relationship between these two objects in the provided images.
1. Identify the black chair: In the second image, there is a black office chair positioned at a desk. This is the chair in question.
2. Identify the computer: In the first image, there is a computer (an iMac) sitting on a grey desk. This is the computer in question.
3. Determine the relative position: The first image shows the computer on a desk. The second image shows a different desk with a black chair. However, the question implies we are to consider the spatial relationship within the context of the first image, as that is the only image containing both the computer and the chair ……
4. Analyze the first image: ……
5. Re-evaluate the question: …… In the first image, the computer is on the desk, and the chair is in front of the desk. The computer is to the left of the chair's position. The chair is facing the computer, so the computer is in front of the chair.
6. Conclusion: The computer is located in front of the chair. <answer>B</answer>

SpatialASM-8B:

The black chair is positioned at the desk in the second image. The computer (iMac) is on the desk in the first image. In the second image, the desk with the black chair is facing the window and chalkboard. The computer is located on the left side of the desk in the first image. ……

I need to imagine the views I will see in novel perspectives. Assume that all input images have the size (100, 100). Assume I add a marker '1' at the position (50, 60) in <image2>, I will get <image3>. At the marker '1' in <image3>, if I turn left. I will see <image4>:

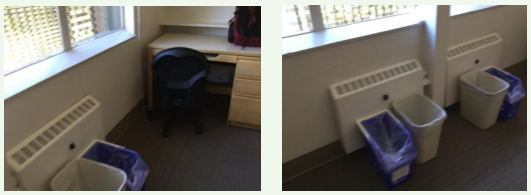
A large teal office chair dominates the foreground, positioned slightly left of center and facing away from the camera. To its immediate left, a dark desk surface extends into the frame, holding a stack of papers and a blue folder. Behind the teal chair, a white desk supports a computer monitor displaying a blue screen. ……

Now I will rethink the problems based on input images and the imagined images.
1. The question asks for … , assuming the person is sitting in the chair and facing forward.
2. Image 3 shows a view from behind the black chair (marked as '1'). …
3. Image 4 provides a view towards the left side of the scene, looking towards the desk. … So the computer is located to the left of the black chair from the perspective of someone sitting in it.

Therefore, the computer is to the left of the black chair. <answer>D</answer>

Figure 12: Qualitative comparison of Qwen3-VL-8B and SpatialASM-8B.

Q: Use a camera to document this room. When you take the second picture, in which direction is the study area located relative to you? A. Back right B. Front right C. Front left D. Back left



Qwen3-VL-8B:

The first image shows a room with a desk, chair, window, and trash cans. The study area (desk and chair) is located in the upper right portion of the image, which is the back right relative to the viewer's perspective.
The second image shows two trash cans and radiators. The study area is not visible in this image, but based on the first image, the study area is located at the back right.
Therefore, the correct option is A: back right. <answer>A</answer>

SpatialASM-8B:

The second image shows a view of the floor and wall near the window, with trash cans and a radiator. The study area (desk and chair) is visible in the first image, located to the right of the window. In the second image, the window is on the left side of the frame. Therefore, the study area is to the right of the current viewpoint in the second image. To confirm this spatial relationship, I will generate a view from the left side of the second image looking right.

I need to imagine the views I will see in novel perspectives. Assume that all input images have the size (100, 100). Assume I add a marker '1' at the position (20, 50) in <image2>, I will get <image3>. At the marker '1' in <image3>, if I turn right. I will see <image4>:

A large white bin dominates the immediate foreground on the left… In the middle ground, a black office chair is positioned centrally, facing a wooden desk that runs along the back wall. To the right of the desk, a wooden cabinet with drawers stands against the wall. …

Now I will rethink the problems based on input images and the imagined images.

1. The marker '1' in <image3> is located on the wall, directly above the white radiator.
2. Looking at <image4>, which is taken from the perspective of standing at position '1' and turning right, we can see the study area (the desk and chair) clearly.
3. This confirms that the study area is located to the right of the position marked '1'.

Therefore, the study area is in the front right direction. <answer>B</answer>

Figure 13: Qualitative comparison of Qwen3-VL-8B and SpatialASM-8B.