BeyondPixels: Internalizing Multi-Faceted Spatial Imagination for Vision-Language Reasoning

Anonymous ACL submission

Abstract

Although recent vision-language models (VLMs) achieve strong results on visual understanding, they continue to fail on tasks that demand precise reasoning about distance, occlusion, layout, and viewpoint. Existing remedies either inject explicit geometric inputs or invoke external tools at inference, both of which externalize spatial competence and leave the model without an internal mechanism for spatial reasoning. We propose **BeyondPixels**, an implicit spatial imagination paradigm in which a VLM, given only RGB inputs, actively generates question-relevant spatial representations at designated latent positions and reasons over them. We instantiate five complementary facets covering local geometry (depth, normal, segmentation), global layout (bird's-eye view), and viewpoint-dependent evidence (novel views), and construct **BeyondPixels-177K** with spatial-tagged reasoning trajectories. A two-stage training procedure first anchors latent positions to geometry-aware targets from a frozen VGGT, then releases them for adaptive task use. Built on Qwen2.5-VL-3B, BeyondPixels reaches 57.3% on VSI-Bench, 51.2% on SQA3D, and 69.3% on SPbench, surpassing GPT-5 and Gemini-3-Pro on VSI-Bench. Ablations and latent visualizations confirm that gains stem from genuine internalization rather than additional data. Code, dataset, and model checkpoints will be publicly released.

1 Introduction

Spatial reasoning is a prerequisite for vision-language models (VLMs) to move toward embodied intelligence, yet it remains a persistent bottleneck. On VSI-Bench (Yang et al., 2025a), even frontier closed-source models such as GPT-5 (OpenAI, 2026) and Gemini-3-Pro (Google DeepMind, 2025) fall well short of human performance. A natural reaction is to attribute this gap to insufficient visual signals. We argue that the underlying cause is more fundamental: a structural mismatch between how VLMs reason and how space must be represented.

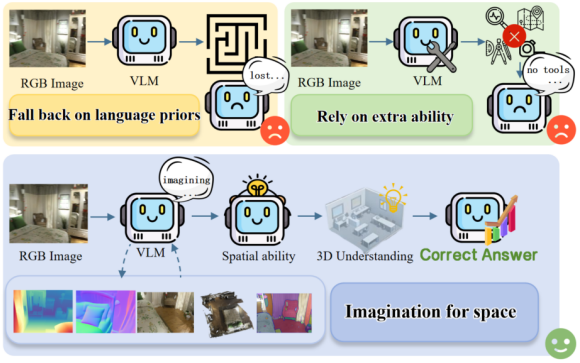


Figure 1: Overview of BeyondPixels. Rather than relying on language priors, 2D shortcuts, or external spatial tools, BeyondPixels internalizes multi-faceted spatial imagination in the latent space, enabling RGB-only geometry-grounded reasoning.

Existing approaches mainly address this issue in two ways. One augments the model with explicit spatial input, such as depth-encoded positional embeddings (Chen et al., 2025a), VGGT-based geometry encoders (Wu et al., 2025b), 3D reconstruction tokens (Fan et al., 2026), camera-guided fusion (Zhao et al., 2025), and persistent memory (Liu et al., 2025b). The other invokes external tools or world models at inference, including novel-view synthesis (Yang et al., 2025b), chain-of-view prompting (Zhao et al., 2026), abstract 3D perception (Liu et al., 2026), and tool-augmented RL (Chen et al., 2025b). Both lines externalize spatial competence: the model itself develops no internal mechanism, and behavior reverts to language priors once auxiliary inputs or external calls are removed. Moreover, they treat spatial evidence as a one-shot injection or a discrete tool call, whereas authentic spatial reasoning requires maintaining and selectively invoking different spatial structures throughout the entire process.

A more promising path is to internalize spatial

knowledge directly into the model's hidden states. Recent attempts confirm this is feasible (Yang et al., 2025c; Chen et al., 2025c; Cao et al., 2025), but they tend to collapse spatial reasoning into a single homogeneous representation or rely on one fixed style of imagined evidence. Human spatial cognition does not work this way. When estimating a room's size, we sketch a rough top-down layout. When judging occlusion, we mentally rotate the view. When ordering objects in depth, we arrange them along a near-to-far axis. The cognitive primitive is not to see more, but to generate the right kind of internal spatial structure on demand. Spatial imagination is multi-faceted and question-conditioned, not a single mental scene.

This observation motivates **BeyondPixels**, an implicit spatial imagination paradigm in which a VLM, given only RGB inputs, actively generates question-relevant spatial representations at designated latent positions and reasons over them. We instantiate five complementary facets that jointly cover local geometry (depth, surface normal, segmentation), global layout (bird's-eye view), and viewpoint-dependent evidence (novel views). To teach this behavior, we construct BeyondPixels-177K, where each sample carries a spatial-tagged reasoning trajectory recording which facets are invoked. Training proceeds in two stages: the first aligns latent positions with offline-extracted spatial targets so that hidden states learn to carry distinct geometric content; the second removes the alignment loss and lets the model autonomously adjust the learned spatial representations to better support downstream QA tasks. Built on Qwen2.5-VL-3B, BeyondPixels reaches 57.3% on VSI-Bench, 51.2% on SQA3D, and 69.3% on SPbench, setting a new state of the art among open-source spatial models and surpassing GPT-5 and Gemini-3-Pro on VSI-Bench.

Our contributions are summarized as follows:

1. We propose **BeyondPixels**, the first framework that internalizes multi-faceted, on-demand spatial imagination into VLM reasoning under an RGB-only inference protocol, distinguishing it from single-representation internalization and from approaches relying on external spatial inputs or external tools.

2. We release **BeyondPixels-177K** together with a spatial-tagged trajectory synthesis pipeline that converts depth, surface normal, segmentation, BEV, and novel-view evidence into structured reasoning supervision.

3. Extensive experiments validate the effectiveness of BeyondPixels, with ablations confirming that gains stem from genuine internalization and visualizations showing that distinct spatial tags induce separable hidden representations.

2 Related Work

Spatial Reasoning in VLMs. Spatial reasoning has emerged as a defining challenge for VLMs aspiring to embodied intelligence (Yang et al., 2025a; Zhang et al., 2025; Ma et al., 2023), and existing solutions fall into two families. The first injects explicit geometric signals into the input or representation pipeline: SD-VLM (Chen et al., 2025a) adds depth-encoded positional embeddings; Spatial-MLLM (Wu et al., 2025b) and VG-LLM (Hu et al., 2025) attach a VGGT (Wang et al., 2025)-based geometry encoder; VLM-3R (Fan et al., 2026) feeds 3D reconstruction tokens; SpaceMind (Zhao et al., 2025) fuses modalities through camera-guided alignment; and VLM^2 (Liu et al., 2025b) maintains persistent memory for long-horizon spatial state. The second treats inference as an interactive process calling external modules: MindJourney (Yang et al., 2025b) couples a VLM with a controllable world model to synthesize new views; CoV (Zhao et al., 2026) performs training-free chain-of-view prompting; Abstract3D (Liu et al., 2026) builds interpretable 3D scene abstractions; SpaceTools (Chen et al., 2025b) learns to invoke depth, segmentation, and pose tools on demand. Both families externalize spatial competence: once auxiliary signals or tool calls are removed, behavior reverts to language priors. BeyondPixels takes the opposite path: we use diverse spatial evidence only as offline supervision and internalize multi-faceted spatial imagination into hidden states, so inference depends on no external scaffold beyond RGB inputs.

Latent and Internalized Reasoning. A separate thread moves intermediate reasoning out of the natural-language surface and into the model's latent space, preserving information that is hard to verbalize. Coconut (Hao et al., 2025) shows that reasoning steps can propagate as continuous hidden states rather than decoded text. Latent Visual Reasoning (Li et al., 2025a) and DMLR (Liu et al.,

2025a) extend this idea to multimodal settings, interleaving latent and explicit thoughts during decoding. MCout (Pham and Ngo, 2025) and Mull-Tokens (Ray et al., 2025) explore modality-agnostic latent thinking, where textual and visual cues share a unified latent space. MVoT (Li et al., 2025b) interleaves textual reasoning with synthesized images; Mirage (Yang et al., 2025c) replaces image generation with latent visual tokens grounded in image embeddings. Closest to our setting, two recent works internalize spatial structure for VLMs. 3DThinker (Chen et al., 2025c) aligns hidden states to a single global 3D representation derived from VGGT, forming a holistic 3D mental scene. SpatialDreamer (Cao et al., 2025) incentivizes active mental imagery through reinforcement learning over a single viewpoint stream. These works confirm internalization is feasible but collapse spatial reasoning into one homogeneous representation. BeyondPixels departs in two ways: spatial imagination is inherently *multi-faceted*, with five facets spanning local geometry, global layout, and viewpoint-dependent evidence; and spatial-tagged trajectories teach the model to invoke facets *on demand*, mirroring how humans switch between mental maps, rotations, and depth orderings.

3 Method

We propose **BeyondPixels**, an implicit spatial imagination framework that enables a VLM to actively generate, maintain, and selectively invoke multi-faceted internal spatial representations during reasoning. As illustrated in Figure 3, the framework consists of three components: a spatial-tagged reasoning corpus that converts heterogeneous auxiliary spatial evidence into structured supervision (Section 3.2), a latent imagination training stage that aligns designated latent positions with geometry-aware target features (Section 3.3), and an adaptive refinement stage that removes external anchoring and allows the model to adapt its internal spatial representations to downstream tasks (Section 3.4). An overview is shown in Figure 3.

3.1 Preliminaries and Notation

We consider a VLM π_θ that takes RGB observations $\mathcal{I} = \{I_i\}_{i=1}^{N}$ and a question Q as input and produces an answer A. Let $H^{(L)} \in \mathbb{R}^{T\times d}$ denote the last-layer hidden states of the generated sequence, where T is the sequence length and d is the hidden dimension.

Spatial facets. We define five spatial facets that span three semantic levels: *local geometry* (depth D, surface normal N, segmentation S), *global layout* (bird's-eye view B), and *viewpoint-dependent evidence* (novel view V). Let $\mathcal{M} = \{\mathtt{D}, \mathtt{N}, \mathtt{S}, \mathtt{B}, \mathtt{V}\}$ denote the facet set.

Spatial tags and latent pads. For each facet $m \in \mathcal{M}$, we introduce paired tag tokens $\langle m_{\text{start}}\rangle$ and $\langle m_{\text{end}}\rangle$. Between them lies a block of K *latent pad tokens* at position set $I_m = \{a_m, \ldots, a_m + K - 1\}$. The hidden states $Z_m = H^{(L)}[I_m] \in \mathbb{R}^{K\times d}$ at these positions are reserved to carry the model's internal imagination of facet m. Unlike text tokens, latent pads are never decoded; they serve as a writable spatial workspace within the language model.

Inference protocol. At inference time, π_θ receives only $(\mathcal{I}, Q)$ and autoregressively generates a spatial-tagged trajectory followed by the final answer. No auxiliary spatial inputs, external generators, or tool calls are used.

3.2 Multi-Faceted Spatial Supervision

To teach the model how and when to invoke each facet, we construct **BeyondPixels-177K**, a corpus in which every sample $(\mathcal{I}, Q, A, \mathcal{F}, \mathcal{R})$ is augmented with auxiliary spatial evidence $\mathcal{F}$ and a spatial-tagged reasoning trajectory $\mathcal{R}$. The full pipeline is shown in Figure 2. We describe the pipeline at a high level here and defer implementation details to Appendix A.

Question-aware keyframe selection. Raw videos and multi-view inputs contain many weakly relevant frames. We use Qwen3.6-Plus (Qwen Team, 2026) to select a compact subset $\mathcal{I}^\star \subseteq \mathcal{I}$ of question-relevant keyframes, on which subsequent spatial evidence is extracted.

Multi-faceted spatial evidence extraction. For each keyframe we extract view-level evidence: depth via Depth Anything V2 (Yang et al., 2024), surface normals via DSINE (Bae and Davison, 2024), and segmentation via SAM 2 (Ravi et al., 2024). At the scene level, we build BEV maps by back-projecting VGGT-predicted (Wang et al., 2025) depth into a shared world frame, followed by ground alignment and rasterization. For samples drawn from SQA3D, we additionally incorporate the human-annotated bird-view representations provided by the dataset as supplementary BEV supervision. For novel views, we

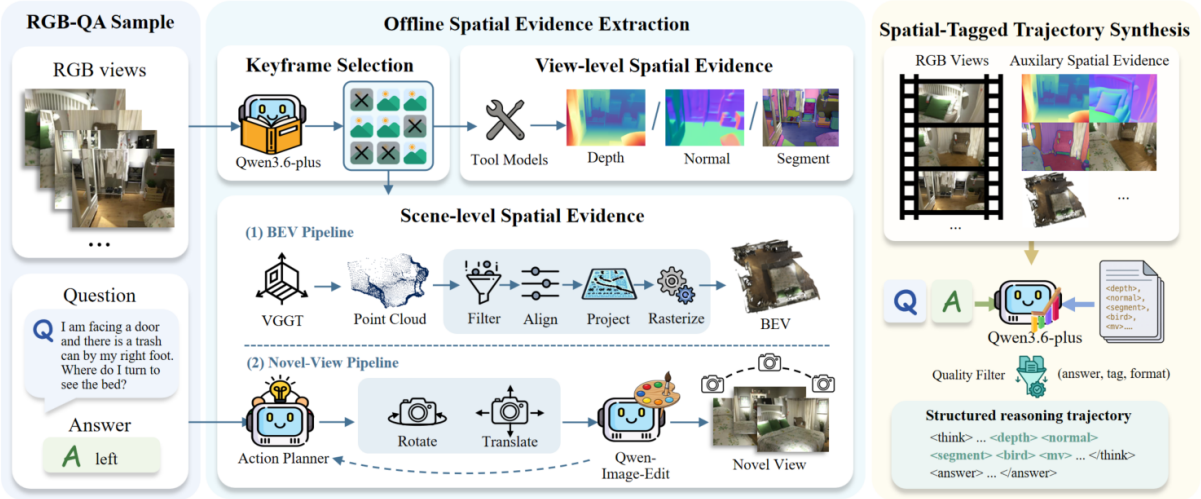

RGB-QA Sample
RGB views
Question
I am facing a door and there is a trash can by my right foot. Where do I turn to see the bed?
Answer
left
Offline Spatial Evidence Extraction
Keyframe Selection
Qwen3.6-plus
View-level Spatial Evidence
Tool Models
Depth
Normal
Segment
Scene-level Spatial Evidence
(1) BEV Pipeline
VGGT
Point Cloud
Filter
Align
Project
Rasterize
BEV
(2) Novel-View Pipeline
Action Planner
Rotate
Translate
Qwen-Image-Edit
Novel View
Spatial-Tagged Trajectory Synthesis
RGB Views
Auxilary Spatial Evidence
Qwen3.6-plus
Quality Filter
(answer, tag, format)
Structured reasoning trajectory
<think> ... <depth> <normal> <bird> <mv> ... </think> <answer> ... </answer>

Figure 2: Construction pipeline of **BeyondPixels-177K**. Per-question keyframes feed view-level and scene-level evidence extraction, after which Qwen3.6-Plus weaves the evidence into spatial-tagged reasoning trajectories.

run an action-conditioned multi-hop exploration: Qwen3.6-Plus plans camera actions conditioned on $(I_0, Q, \text{history})$, and Qwen-Image-Edit (Wu et al., 2025a) synthesizes the resulting view, until sufficient evidence is gathered or a hop budget is reached.

Trajectory synthesis. Given $\mathcal{F}$, Qwen3.6-Plus identifies the subset of facets required by Q, selects the most informative evidence per facet, and weaves them into a spatial-tagged reasoning trajectory $\mathcal{R}$ rather than indiscriminately injecting all evidence. Each trajectory has the schema:

$$\mathcal{R}: \quad \big[\,\text{prose}_i \;\; \langle m_i\rangle \cdots \langle/m_i\rangle\,\big]_{i=1}^{n} \;\; \langle\text{ans}\rangle\, A\, \langle/\text{ans}\rangle, \tag{1}$$

where each $\langle m_i\rangle \cdots \langle/m_i\rangle$ block contains the K latent pad tokens reserved for facet m_i. Low-quality trajectories are filtered by answer consistency, tag-evidence matching, and format validation. The final corpus contains 177K samples; statistics and source breakdown are in Appendix B.

The key design choice here is that $\mathcal{R}$ encodes both *which* facets to invoke and *where* to invoke them. This structure provides the supervision signal needed for the next stage.

3.3 Stage 1: Latent Imagination Training

The goal of Stage 1 is to make the latent pad positions carry geometrically meaningful content, so that hidden states themselves encode spatial structure rather than only text.

Text objective. We supervise the generation of all non-latent tokens (prose, tag tokens, and the final answer) with standard cross-entropy:

$$\mathcal{L}_{\text{text}} = -\sum_{t \in \mathcal{M}_{\text{text}}} \log \pi_\theta(y_t \mid y_{<t}, \mathcal{I}, Q), \tag{2}$$

where $\mathcal{M}_{\text{text}}$ is the set of text-token positions in $\mathcal{R}$. Latent pad positions are masked out from this loss.

Latent spatial alignment. For each facet m active in a sample, we extract its latent states $Z_m = H^{(L)}[I_m]$ and project them into the target spatial feature space via a facet-specific projector P_m:

$$\hat{F}_m = P_m(Z_m) \in \mathbb{R}^{K \times d_F}. \tag{3}$$

The target feature F_m^{VGGT} is extracted offline by a frozen VGGT encoder applied to the corresponding spatial evidence in $\mathcal{F}$. We align $\hat{F}_m$ to F_m^{VGGT} with a Frobenius loss:

$$\mathcal{L}_{\text{imagine}} = \frac{1}{|\mathcal{S}|} \sum_{m \in \mathcal{S}} \big\| \hat{F}_m - F_m^{\text{VGGT}} \big\|_F^2, \tag{4}$$

where $\mathcal{S} \subseteq \mathcal{M}$ is the set of facets supervised in the current sample. Targets enter only as training-time alignment signals and are never seen at inference.

Stage 1 objective. The full loss combines both terms:

$$\mathcal{L}^{(1)} = \lambda_{\text{text}} \mathcal{L}_{\text{text}} + \lambda_{\text{imagine}} \mathcal{L}_{\text{imagine}}. \tag{5}$$

Through Stage 1, each tag-bounded latent block becomes a geometry-aware slot: depth pads learn to encode front-back ordering and metric distance, BEV pads learn to encode top-down layout, and so on.

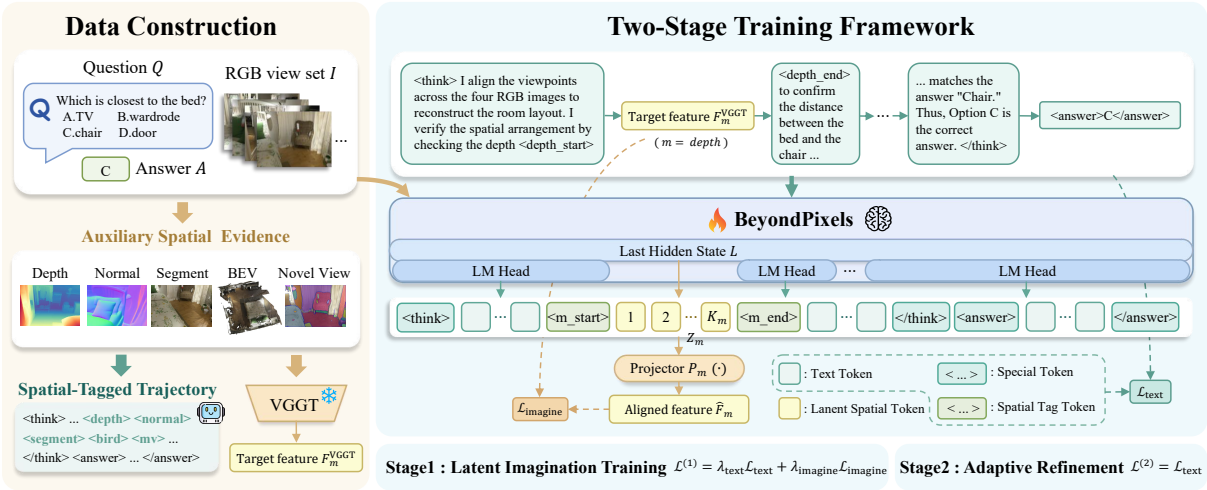

Data Construction
Question Q
RGB view set I
Which is closest to the bed?
A.TV B.wardrode
C.chair D.door
C
Answer A
Auxiliary Spatial Evidence
Depth
Normal
Segment
BEV
Novel View
Spatial-Tagged Trajectory
<think> ... <depth> <normal> <bird> <mv> ... </think> <answer> ... </answer>
VGGT
Target feature F_m^{VGGT}
Two-Stage Training Framework
<think> I align the viewpoints across the four RGB images to reconstruct the room layout. I verify the spatial arrangement by checking the depth <depth_start>
Target feature F_m^{VGGT}
(m = depth)
<depth_end> to confirm the distance between the bed and the chair ...
... matches the answer "Chair." Thus, Option C is the correct answer. </think>
<answer>C</answer>
BeyondPixels
Last Hidden State L
LM Head
LM Head
LM Head
<think>
<m_start>
1
2
K_m
<m_end>
</think>
<answer>
</answer>
Z_m
Projector P_m (·)
Aligned feature $\hat{F}_m$
$\mathcal{L}_{imagine}$
: Text Token
: Special Token
: Lanent Spatial Token
: Spatial Tag Token
$\mathcal{L}_{text}$
Stage1 : Latent Imagination Training $\mathcal{L}^{(1)} = \lambda_{text}\mathcal{L}_{text} + \lambda_{imagine}\mathcal{L}_{imagine}$
Stage2 : Adaptive Refinement $\mathcal{L}^{(2)} = \mathcal{L}_{text}$

Figure 3: Overview of **BeyondPixels**. The model reserves latent pad positions for five spatial facets and is trained in two stages: Stage 1 aligns the pads with geometry-aware targets extracted offline by a frozen VGGT, and Stage 2 removes the alignment loss so that the model learns to autonomously invoke the right facet during RGB-only inference. Auxiliary spatial evidence is used only as training-time supervision and never appears at inference.

3.4 Stage 2: Adaptive Refinement

Stage 1 forces latent pads to carry specific geometric content, but it does not yet teach the model to *use* this content for downstream reasoning. If we kept the alignment loss throughout training, latent pads would remain tightly anchored to VGGT-shaped targets and could not be flexibly recombined with text reasoning. Stage 2 therefore removes the alignment loss and lets the latent representations adapt to the task:

$$\mathcal{L}^{(2)} = \mathcal{L}_{\text{text}}. \tag{6}$$

Critically, the model still emits spatial tags and still writes into latent pads at the corresponding positions, but it now learns autonomously when to open a facet block, how long to dwell on it, and how to translate latent content into answer-relevant prose. This stage bridges the gap between *forming* an internal imagination (Stage 1) and *deploying* it within an RGB-only reasoning trajectory.

4 Experiments

We evaluate BeyondPixels along three axes: (i) whether it achieves strong overall spatial reasoning across diverse benchmarks, (ii) whether the observed gains genuinely come from internalized spatial imagination rather than data scale or auxiliary inputs, and (iii) what design choices shape the internal spatial representations.

4.1 Experimental Setup

Implementation. We build BeyondPixels on Qwen2.5-VL-3B (Bai et al., 2025) and train on BeyondPixels-177K for one epoch with AdamW, batch size 64, and learning rate 1×10^{-5}. Stage 1 additionally minimizes $\mathcal{L}_{\text{imagine}}$ with $\lambda_{\text{imagine}} = 0.1$. All experiments use 4 NVIDIA RTX PRO 6000 Blackwell GPUs. Full hyperparameters are in Appendix C.

Benchmarks. We evaluate on VSI-Bench (Yang et al., 2025a), SQA3D (Ma et al., 2023), SPAR-bench (Zhang et al., 2025), SPbench (Li et al., 2025c), and CV-Bench (Tong et al., 2024), spanning video, multi-view, and single-view settings. Numerical VSI sub-tasks use MRA; others use accuracy. All evaluations are zero-shot. Details are in Appendix D.

Baselines. We compare with three families: *closed-source VLMs* (GPT-4o (OpenAI, 2024), Gemini-2.0-Flash (Comanici et al., 2025), Gemini-3-Pro (Google DeepMind, 2025), GPT-5 (OpenAI, 2026)); *open-source general VLMs* (Qwen2.5-VL-3B/7B (Bai et al., 2025), InternVL-2.5-4B/8B (Chen et al., 2024), LLaVA-OneVision-7B (Li et al., 2024)); and *open-source spatial models* covering three paradigms: spatial augmentation (SpatialLadder-3B (Li et al., 2025c), Spatial-MLLM-4B (Wu et al., 2025b)), RL for spatial reasoning (SpaceR-7B (Ouyang et al., 2025), ViLaSR-7B (Wu et al., 2025c)), and internalized 3D imagination (3DThinker-3B (Chen et al., 2025c)).

Table 1: Comparison on VSI-Bench (Yang et al., 2025a). Numerical-answer tasks are evaluated with MRA, while multiple-choice-answer tasks are evaluated with accuracy. Best results are highlighted in **bold**.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Human	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
Random Choice (Freq.)	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Closed Models									
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-2.0-Flash	45.4	56.2	30.9	66.7	31.8	51.3	46.3	24.5	55.1
Gemini-3-Pro	52.5	38.0	37.8	72.7	44.1	59.9	55.7	45.9	66.0
GPT-5	55.0	53.3	34.5	73.3	47.5	63.7	48.7	50.3	68.9
Open-source General Models									
Qwen2.5-VL-3B	27.6	31.0	20.1	16.5	25.3	38.1	44.1	25.2	20.4
Qwen2.5-VL-7B	32.3	32.9	18.2	43.9	31.7	38.0	37.4	28.4	28.0
InternVL-2.5-4B	32.6	45.0	15.5	37.5	24.6	37.2	41.5	31.4	26.2
InternVL-2.5-8B	40.2	50.6	31.1	40.2	39.3	45.1	41.4	29.4	43.9
LLaVA-OneVision-7B	33.1	46.1	26.2	36.3	29.5	30.8	37.2	35.1	21.8
Open-source Spatial Intelligence Models									
SpatialLadder-3B	44.9	62.2	35.4	62.0	41.4	45.6	46.5	27.3	38.5
Spatial-MLLM-4B	46.3	66.7	38.1	63.6	35.5	40.4	48.2	33.0	44.3
SpaceR-7B	41.5	44.5	24.7	53.5	37.3	41.9	46.1	29.3	**54.8**
3Dthinker-3B	53.2	–	–	–	–	–	–	–	–
ViLaSR-7B	44.6	58.1	33.9	61.4	28.9	45.1	46.5	29.9	53.2
Ours									
BeyondPixels-3B	**57.3**	61.7	**41.8**	**72.1**	**64.3**	**57.8**	48.0	**41.2**	**70.9**

Table 2: Comparison with existing spatial reasoning models on SQA3D (Ma et al., 2023), SPARbench (Zhang et al., 2025), SPbench (Li et al., 2025c), and CVbench (Tong et al., 2024). **Bold** numbers indicate the best performance, while underline numbers represent the second-best performance.

Model	SQA3D	SPARbench				SPbench			CVbench			Overall
	EM	Low	Mid	High	Avg.	SI	MV	Avg.	2D	3D	Avg.	
ViLaSR-7B	41.0	**33.7**	41.0	42.5	37.6	50.2	57.6	53.9	**75.3**	76.5	**75.9**	52.1
SpatialLadder-3B	36.8	25.8	33.2	42.6	34.4	**70.2**	**70.9**	**70.6**	72.4	74.9	73.7	53.9
Spatial-MLLM-4B	**55.9**	29.9	20.3	38.1	32.2	43.7	53.1	48.4	–	–	73.8	52.6
SpaceR-7B	42.5	–	–	–	35.9	48.6	59.4	54.0	73.6	**76.9**	75.3	52.0
Qwen2.5-VL-3B	39.3	15.3	26.4	32.2	24.6	40.3	36.6	38.5	69.1	72.2	70.6	43.3
BeyondPixels-3B (ours)	51.2	30.8	**42.7**	**51.6**	**39.1**	70.1	68.4	69.3	72.4	76.0	74.2	**58.4**

4.2 Main Results

Fine-grained spatial reasoning on VSI-Bench. Table 1 reports per-task results on VSI-Bench. BeyondPixels-3B reaches **57.3%** overall, lifting its Qwen2.5-VL-3B backbone by **+29.7** absolute points (27.6 → 57.3) and outperforming all open-source spatial baselines. With only 3B parameters, it also surpasses every 7B–8B open-source model by at least 17 points, and exceeds the strongest closed-source models GPT-5 (55.0%) and Gemini-3-Pro (52.5%), suggesting that internalized spatial imagination is more parameter-efficient than simply scaling model size. The gains concentrate on tasks that require explicit geometric grounding: Absolute Distance (**41.8%**, vs. GPT-4o at 5.3% and GPT-5 at 34.5%), Object Size (**72.1%**), Room Size (**64.3%**, +16.8 over GPT-5), and Relative Distance (**57.8%**). BeyondPixels also leads Route Planning (**41.2%**) and reaches **70.9%** on Appearance Order, the latter being a temporal task that benefits indirectly from BEV and novel-view supervision through stronger scene-level grounding. Remaining gaps on Object Count and Relative Direction suggest that fine-grained instance disambiguation and reference-frame normalization are not yet addressed by our current facet set, pointing to natural directions for extending the facet design in future work. Compared with 3DThinker, which uses a single holistic 3D representation, BeyondPixels improves overall by +4.1 points, supporting our central hypothesis that multi-faceted, on-demand imagination is more effective than a single homogeneous mental scene.

Table 3: Ablation study on training strategies and input settings. *Aux* denotes directly feeding auxiliary spatial representations as additional inputs. Best results are highlighted in bold.

Setting	VSIbench	SPARbench	CVbench
Qwen2.5-VL-3B	27.6	24.6	70.3
Qwen2.5-VL-3B + *Aux*	27.9	24.8	68.3
Train with QA	55.7	36.7	64.7
Train with QA + *Aux*	55.8	36.9	64.3
w/o stage1	55.3	36.3	71.6
w/o stage2	55.1	37.4	74.1
Ours	**57.3**	**39.1**	**74.2**

Generalization across spatial benchmarks. Table 2 evaluates the same model on four additional benchmarks. BeyondPixels-3B attains the best overall score of **58.4%**, ahead of SpatialLadder-3B (53.9%), Spatial-MLLM-4B (52.6%), ViLaSR-7B (52.1%), and SpaceR-7B (52.0%). On SPARbench, the advantage widens with task difficulty (Low: 30.8 vs. ViLaSR's 33.7; Mid: **42.7**, +1.7; High: **51.6**, +9.0 over SpatialLadder), indicating that internalized imagination is most beneficial for compositional spatial reasoning rather than primitive perception. On SQA3D, BeyondPixels ranks second (51.2%), behind Spatial-MLLM-4B (55.9%) which is purpose-built with dense annotation tailored to SQA3D's format; we attain comparable performance without such task-specific scaffolding, while remaining +14.4 points above SpatialLadder-3B. On SPbench, BeyondPixels matches SpatialLadder on single-image reasoning (70.1 vs. 70.2) while being substantially more general across the broader benchmark suite. On CV-Bench, BeyondPixels reaches 74.2%, close to the top baselines (75.9% at best), indicating that spatial specialization does not sacrifice general visual understanding. Together with VSI-Bench, the consistent gains across five benchmarks indicate that internalized multi-faceted imagination generalizes beyond the training distribution rather than overfitting to a single benchmark.

4.3 Ablation Study

Auxiliary inputs alone are not enough, and can be harmful. Feeding depth, normal, segmentation, and BEV images directly to Qwen2.5-VL-3B at inference (*Qwen2.5-VL-3B + Aux*) yields negligible change on VSI-Bench (27.6% → 27.9%) and *degrades* CV-Bench by 2.0% (70.3% → 68.3%, Table 3). Without supervision, non-RGB distributions disrupt pretrained visual priors.

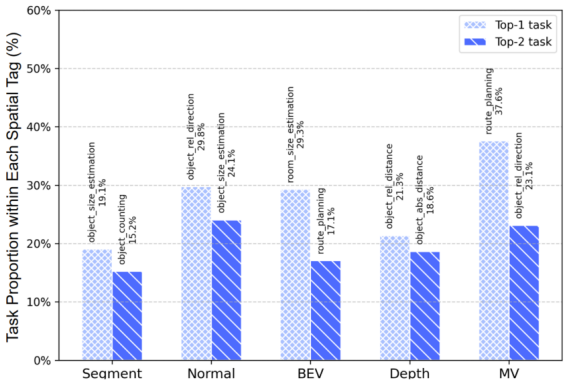


Figure 4: Task distribution within each spatial tag. Top-1 and Top-2 bars denote the most frequent and second most frequent task types associated with each auxiliary spatial representation.

Table 4: Ablation study on different spatial supervision designs.

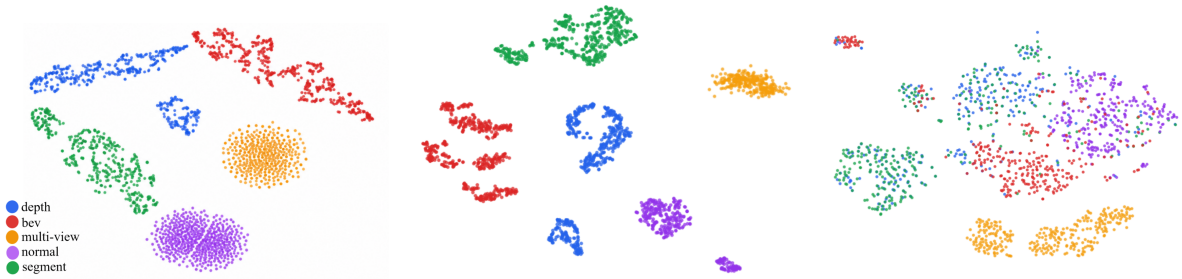
Setting	VSIbench	SPARbench	CVbench
VGGT feature	**57.3**	**39.1**	74.2
fused feature	54.9	36.9	**74.6**

QA-only training falls short, and is unrescued by extra inputs. Training on BeyondPixels-177K with only the QA loss (*Train with QA*) lifts VSI to 55.7% but drops CV-Bench to 64.7%. Adding auxiliary inputs at inference (*Train with QA + Aux*) yields no meaningful difference (within 0.4% on all benchmarks). The bottleneck is not exposure to evidence but the lack of an internal mechanism to organize it.

Both stages contribute, asymmetrically. Removing Stage 1 (*w/o stage1*) drops VSI by 2.0% and CV by 2.6%, while removing Stage 2 (*w/o stage2*) drops VSI by 2.2% but leaves CV nearly unchanged. Stage 1 installs geometry-aware structure that protects general visual ability; Stage 2 teaches the model to deploy it for spatial reasoning. Neither alone reproduces the full effect.

4.4 Analyzing the Internal Representations

Facets are invoked on demand. A central claim of our design is that spatial imagination should be question-conditioned: different tasks should activate different facets. Figure 4 reports, for each facet, the two most frequent VSI-Bench task types in which the facet is invoked. The pattern aligns with geometric intuition: depth for distance estimation, BEV for room-size and route planning, normal for relative direction and size, segment



(a) BeyondPixels-stage2 (b) BeyondPixels-Stage1 (c) Qwen2.5-VL-3B

Figure 5: Visualization of spatial auxiliary representations in the latent space. Compared with Qwen2.5-VL-3B, BeyondPixels shows clearer separation among different spatial representations.

Table 5: Ablation study on the effect of latent token size.

Settings	VSIbench	SPARbench	CVbench
Latent size 4	56.7	38.2	72.1
Latent size 8	**57.3**	**39.1**	**74.2**
Latent size 12	56.8	38.0	73.0
Latent size 16	53.6	36.6	71.3

for boundary-sensitive tasks (size, counting), and novel-view for multi-viewpoint reasoning (route planning, relative direction). No facet collapses to a single task or is uniformly activated. This task-conditioned routing emerges purely from spatial-tagged trajectory supervision, with no explicit task-facet binding, and directly validates the on-demand invocation behavior that motivates our paradigm.

Pure 3D targets outperform fused 2D/3D targets. Besides VGGT 3D features, we follow Spatial-MLLM's dual-encoder design and construct fused 2D/3D features via a connector pre-trained on BeyondPixels-177K. Table 4 shows that pure VGGT gives a clear advantage on spatial benchmarks (VSI: $+2.4\%$, SPAR: $+2.2\%$), while the two are comparable on CV-Bench (within 0.4%). Fused targets mix appearance and semantics, diluting the depth, layout, and viewpoint structure we want latent pads to encode.

Latent capacity needs a sweet spot. Table 5 varies the latent pad size K. Accuracy improves modestly from $K = 4$ to $K = 8$ (-0.6% on VSI when reduced) but degrades sharply when K exceeds the optimum ($K = 16$ loses 3.7% on VSI). The asymmetry indicates that the main cost of mis-sized capacity is optimization difficulty, not information bottleneck. We use $K = 8$ throughout.

Latent tokens form facet-specific clusters. We visualize latent pad hidden states with t-SNE and further compare the latent organization of BeyondPixels-stage-2 with the base Qwen2.5-VL-3B model and BeyondPixels-stage-1 in Figure 5. BeyondPixels-Stage1 already produces distinguishable clusters for different spatial facets, showing that the alignment objective successfully injects facet-specific geometric information into the hidden states. Besides, tokens from different spatial tags form clearly separable clusters, indicating that BeyondPixels-stage2 learns facet-specific internal representations rather than a single shared spatial code. The depth cluster further splits into two sub-clusters, consistent with a disentanglement between relative depth (front-back ordering) and scene-scale cues (absolute distance, spatial extent). In contrast, Qwen2.5-VL-3B shows fragmented and entangled clusters, where hidden states from different tags are heavily mixed. Additional cases are in Appendix F. This emergent structure aligns with our multi-faceted hypothesis: spatial imagination is a set of complementary, question-conditioned representations.

5 Conclusion

We presented **BeyondPixels**, an implicit spatial imagination paradigm that internalizes multi-faceted spatial reasoning into a VLM under an RGB-only inference protocol. By aligning latent pad positions to depth, normal, segmentation, BEV, and novel-view targets in Stage 1 and releasing them for adaptive use in Stage 2, the model learns to invoke task-relevant spatial structures on demand. Trained on BeyondPixels-177K, BeyondPixels-3B sets a new state of the art on VSI-Bench, SQA3D, and SPbench, surpassing GPT-5 and Gemini-3-Pro on VSI-Bench.

Limitations

BeyondPixels has several limitations. First, our spatial supervision relies on offline-constructed auxiliary evidence whose quality bounds model performance: BEV reconstruction depends on monocular depth and camera pose estimation, and novel-view synthesis may introduce geometric inconsistencies. Second, latent-space visualizations show that distinct facets induce separable hidden representations, but the precise semantic structure within each facet is not yet fully interpretable. Third, our facet set covers local geometry, global layout, and viewpoint-dependent evidence, but does not directly address fine-grained instance disambiguation or reference-frame normalization, which remain weak spots on Object Count and Relative Direction. Future work may explore controllable mechanisms for internal spatial imagination, more comprehensive facet designs, and tighter coupling between latent imagination and answer generation.

Ethics Statement

This work focuses on improving spatial reasoning in vision-language models using publicly available datasets (ScanNet-derived corpora, SQA3D, SPAR-7M, SATv2, SpatialLadder) and does not collect new human data or involve human subjects. All source datasets are used in accordance with their original licenses and intended research purposes. Our auxiliary evidence is generated entirely from RGB observations through off-the-shelf perception models, and no personal, biometric, or sensitive information is introduced. During data construction, we use large language models (Qwen3.6-Plus) to assist with keyframe selection, camera-action planning, and spatial-tagged trajectory synthesis, and Qwen-Image-Edit to synthesize novel views; all such outputs are used as training-time supervision only, do not appear at inference, and are filtered for quality and consistency with ground-truth answers before being added to the corpus. We also use large language models to polish the writing of this manuscript; all technical content, claims, and experimental results are produced and verified by the authors. Although improved spatial reasoning has positive applications in robotics, accessibility, and embodied AI, we acknowledge that spatial understanding can in principle be misused in surveillance contexts. We release BeyondPixels-177K and trained checkpoints for research use, and encourage downstream practitioners to follow established responsible-AI practices.

References

Gwangbin Bae and Andrew J. Davison. 2024. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. 2025a. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *Preprint*, arXiv:2509.17664.

Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. 2025b. Spacetools: Tool-augmented spatial reasoning via double interactive rl. *Preprint*, arXiv:2512.04069.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 2025c. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. 2026. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *Preprint*, arXiv:2505.20279.

Google DeepMind. 2025. Gemini 3 pro: Model evaluation – approach, methodology & results.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. 2025. G ^{2} vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025a. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulic, and Furu Wei. 2025b. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

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

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. 2025a. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*.

Yifan Liu, Fangneng Zhan, Kaichen Zhou, Yilun Du, Paul Pu Liang, and Hanspeter Pfister. 2026. Abstract 3d perception for spatial intelligence in vision-language models. *Preprint*, arXiv:2511.10946.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. 2025b. Vision-language memory for spatial reasoning. *Preprint*, arXiv:2511.20644.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*.

OpenAI. 2024. Gpt-4o system card.

OpenAI. 2026. Gpt-5 system card. *Preprint*, arXiv:2601.03267.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*.

Tan-Hanh Pham and Chris Ngo. 2025. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*.

Qwen Team. 2026. Qwen3.6-Plus: Towards real world agents.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. 2025. Mull-tokens: Modality-agnostic latent thinking. *arXiv preprint arXiv:2512.10941*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Aditya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2025b. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*.

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

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *arXiv:2406.09414*.

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

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025c. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 2025. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *Advances in Neural Information Processing Systems*.

Haoyu Zhao, Akide Liu, Zeyu Zhang, Weijie Wang, Feng Chen, Ruihan Zhu, Gholamreza Haffari, and Bohan Zhuang. 2026. Cov: Chain-of-view prompting for spatial reasoning. *Preprint*, arXiv:2601.05172.

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

A Dataset Construction Details

This appendix expands the data construction pipeline introduced in Section 3.2. The pipeline operates entirely offline and produces, for each RGB-QA sample, a set of auxiliary spatial evidence $\mathcal{F}$ and a spatial-tagged reasoning trajectory $\mathcal{R}$. None of these signals enter the model at inference time.

A.1 Source Datasets

We construct BeyondPixels-177K by aggregating RGB-QA pairs from six complementary spatial reasoning datasets and benchmarks. These sources cover egocentric video, embodied 3D scene understanding, synthetic spatial aptitude tasks, multi-view spatial perception, and 3D reconstructive instruction following.

- **VSI-590K & VSI10K**: large-scale image and video QA from Cambrian-S, focusing on visual-spatial reasoning over object count, bounding boxes, room size, distance, and spatial relations.
- **SQA3D**: a situated question-answering dataset for 3D scenes. Given a textual description of the agent's position and orientation, the model reasons about the surrounding environment from that situated viewpoint. Built on 650 ScanNet scenes, with 6.8K situations, 20.4K descriptions, and 33.4K reasoning questions.
- **VLM-3R**: a video QA dataset for 3D spatial reasoning, with approximately 340K training samples drawn from indoor 3D scene data including ScanNet. Tasks include object localization, navigation, route planning, and scene layout reasoning.
- **SATv2**: a spatial reasoning training set built from 3D simulated environments, with $\sim$175K QA pairs and 20K scenes, covering static relations, camera motion, object movement, action consequence, and perspective transformation.
- **SpatialLadder**: a multimodal spatial reasoning corpus for progressive training, with 26,610 samples spanning object localization, single-image, multi-view, and video spatial reasoning.
- **SPAR-Bench**: a spatial perception and reasoning benchmark built from SPAR-7M with human verification, containing 7,207 high-quality QA pairs across single-view, multi-view, and video settings.

The distribution of training samples across sources is shown in Figure 6.

A.2 Question-Aware Keyframe Selection

For video and multi-view samples, the raw observation sequence often contains redundant or weakly relevant frames. We use Qwen3.6-Plus (Qwen Team, 2026) as a question-aware selector. Given the input frames $\mathcal{I}$ and question Q, the selector returns a compact subset $\mathcal{I}^\star \subseteq \mathcal{I}$ scored by spatial informativeness (e.g., object location, depth ordering, occlusion, layout, viewpoint change). The selection prompt is given in Figure 13. Frames are scored individually and aggregated by importance; the final set contains the minimal complementary subset rather than a fixed-size cut.

A.3 Multi-Faceted Spatial Evidence Extraction

At the view level, we extract depth via Depth Anything V2 (Yang et al., 2024), surface normals via DSINE (Bae and Davison, 2024), and segmentation via SAM 2 (Ravi et al., 2024). All three operate

on the keyframe subset $\mathcal{I}^\star$. Examples are shown in Figure 7.

At the scene level, we construct two complementary representations:

BEV maps. For each keyframe set we use VGGT (Wang et al., 2025) to estimate camera parameters and per-pixel depth, back-project valid pixels into a shared world coordinate system, filter outliers, align to the ground plane via RANSAC, and rasterize the resulting point cloud into a top-down BEV grid. For samples drawn from SQA3D, we additionally incorporate the human-annotated bird-view representations provided by the dataset as supplementary supervision. BEV examples are in Figure 8.

Novel-view observations. We run an action-conditioned multi-hop exploration loop. Starting from an initial view I_0, Qwen3.6-Plus plans a camera action conditioned on $(I_0, Q, \text{history})$. The action space contains rotate (left/right/up/down) and translate (left/right/forward/backward); rotate uses a degree parameter, translate uses a distance in meters. Each structured action is converted to a textual prompt (Figures 15–16) and passed to Qwen-Image-Edit (Wu et al., 2025a) (Figure 17) to synthesize the new viewpoint. The new view is appended to the observation history and the loop continues until either the model commits to a final answer or a hop budget is exhausted. Novel-view examples are in Figure 9.

A.4 Spatial-Tagged Trajectory Synthesis

Given the auxiliary evidence $\mathcal{F}$, Qwen3.6-Plus identifies the subset of facets relevant to Q from $\mathcal{M} = \{\texttt{depth}, \texttt{normal}, \texttt{segment}, \texttt{bird}, \texttt{mv}\}$, selects the most informative evidence per facet, and weaves them into a structured reasoning trajectory $\mathcal{R}$. The trajectory is wrapped in `<think>...</think>` blocks, with each facet appearing between paired tags (e.g., `<depth_start>...<depth_end>`). The final answer is wrapped in `<answer>...</answer>`. The full synthesis prompt is given in Figure 18.

A.5 Quality Filtering

We filter low-quality trajectories on three criteria: (i) *answer consistency*: the trajectory must conclude with the ground-truth answer; (ii) *tag-evidence matching*: each spatial tag must be supported by visible evidence rather than hallucinated content; (iii) *format validation*: the trajectory must conform to the schema in Equation (1), with no unclosed tags or malformed tokens. Trajectories failing any criterion are discarded. The filtering prompt is given in Figure 14.

B Dataset Statistics and Source Composition

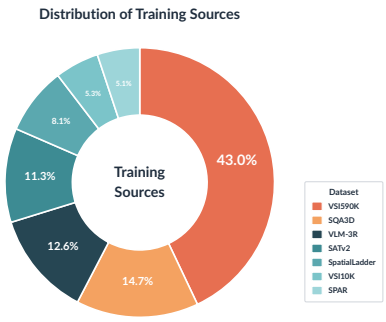


Figure 6: Distribution of training samples across source datasets in BeyondPixels-177K.

The final corpus contains 177K samples after filtering. Figure 6 shows the contribution of each source dataset. Facet activation statistics across question categories will be released with the dataset.

C Training Details

Backbone and infrastructure. We initialize from Qwen2.5-VL-3B (Bai et al., 2025) and train with the bf16 precision on 4 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. The training framework is built on the standard HuggingFace stack with FlashAttention-2 and gradient checkpointing.

Stage 1 hyperparameters. Stage 1 jointly optimizes $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{imagine}}$ over the full BeyondPixels-177K corpus. Key hyperparameters are listed in Table 6. We use AdamW with a cosine learning-rate schedule and a warmup ratio of 3%.

Stage 2 hyperparameters. Stage 2 removes $\mathcal{L}_{\text{imagine}}$ and continues training under $\mathcal{L}_{\text{text}}$ only. All other hyperparameters are identical to Stage 1 (Table 7), so that any performance change is attributable to the change in objective rather than to training configuration drift.

D Evaluation Details

Inference protocol. At inference, BeyondPixels receives only RGB observations and the question.

Table 6: Stage 1 training hyperparameters.

Hyperparameter	Value
per_device_train_batch_size	4
gradient_accumulation_steps	6
global batch size	64
learning rate	1×10^{-5}
optimizer	AdamW
lr_scheduler	cosine
warmup ratio	0.03
num_train_epochs	1
precision	bf16
attention backend	flash_attention_2
gradient checkpointing	true
max_pixels	112896
min_pixels	12544
λ_{text}	0.1
λ_{imagine}	0.1
latent pad size K	8

Table 7: Stage 2 training hyperparameters.

Hyperparameter	Value
per_device_train_batch_size	4
gradient_accumulation_steps	6
global batch size	64
learning rate	1×10^{-5}
optimizer	AdamW
lr_scheduler	cosine
warmup ratio	0.03
num_train_epochs	1
precision	bf16
attention backend	flash_attention_2
gradient checkpointing	true
max_pixels	112896
min_pixels	12544
latent pad size K	8

No auxiliary spatial maps, point clouds, BEV grids, synthesized views, or external tools are used. The model autoregressively generates a spatial-tagged reasoning trajectory followed by the final answer. We use greedy decoding for VSI-Bench (matching the protocol of Yang et al. (2025a)) and nucleus sampling (p=0.9, T=0.7) elsewhere.

Benchmark protocols. For VSI-Bench we follow the official prompt format and use 32 uniformly sampled frames. Numerical sub-tasks (Object Count, Absolute Distance, Object Size, Room Size) are scored with MRA as defined in Yang et al. (2025a); multiple-choice sub-tasks (Relative Distance, Relative Direction, Route Plan, Appearance Order) are scored by accuracy. For SQA3D and SPbench, we use the official train/test splits and report exact-match accuracy. For SPARbench we follow the Low/Mid/High difficulty stratification of the official release. For CV-Bench we report accuracy on the 2D and 3D subsets.

Baselines. Closed-source baselines (GPT-4o, Gemini-2.0-Flash, Gemini-3-Pro, GPT-5) are queried via their public APIs with default temperature and 32 uniformly sampled frames. Open-source baselines are evaluated using the checkpoints released by their authors, with the configurations reported in their respective papers. Numbers marked with $*$ in Table 1 are taken from the corresponding original publications.

E Auxiliary Spatial Representation Examples

We provide qualitative examples of each spatial facet to illustrate the offline supervision signal. These representations are used only during training and do not appear at inference.

Local geometry. Figure 7 shows depth, normal, and segmentation for representative indoor scenes. Depth provides front-back ordering and relative distance; normal captures local surface orientation; segmentation delineates object regions and boundaries. The three together cover the geometric primitives needed for reasoning about distance, occlusion, and region ownership.

Global layout. Figure 8 shows BEV maps obtained by back-projecting VGGT-predicted depth into a shared world frame. The top-down view reorganizes the scene to expose room structure, furniture arrangement, navigable space, and route-level relations.

Viewpoint-dependent evidence. Figure 9 shows novel-view observations generated from planned camera actions. These views expose occluded regions, verify object relations from alternative angles, and support reasoning about visibility and viewpoint change.

F Case Study

We present three qualitative cases that illustrate how BeyondPixels routes spatial evidence through latent pads at inference. In each case, the model autonomously decides which facets to invoke and weaves them into a coherent reasoning trajectory.

These cases also demonstrate that the spatial-tagged reasoning trajectories produced by BeyondPixels explicitly tie each conclusion to a specific facet of internal imagination, rather than relying on unsupported textual chains.

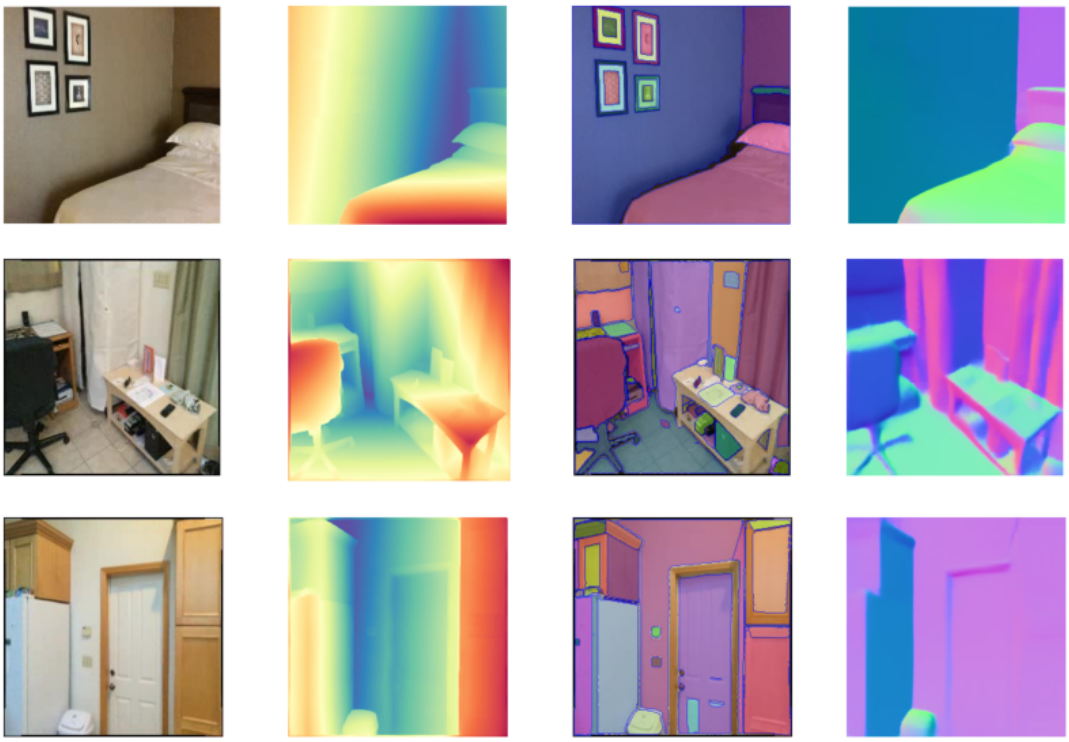


Figure 7: RGB views with corresponding depth, surface normal, and segmentation maps.

G Prompt Templates

We provide the prompt templates used in the data construction pipeline (Section A). All prompts use structured input fields and constrained output formats to enable automatic processing and downstream filtering.

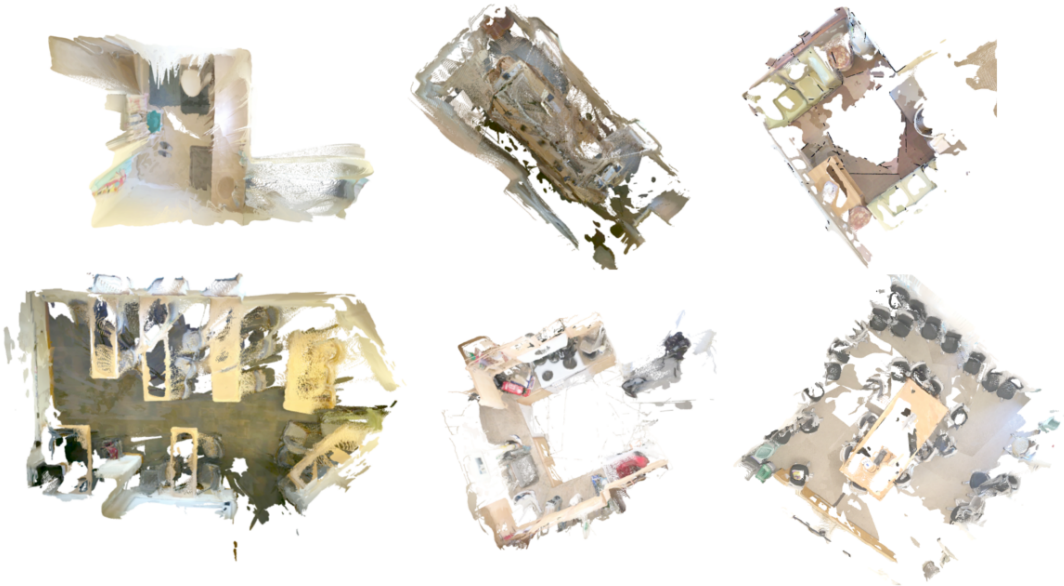


Figure 8: BEV representations for representative indoor scenes.

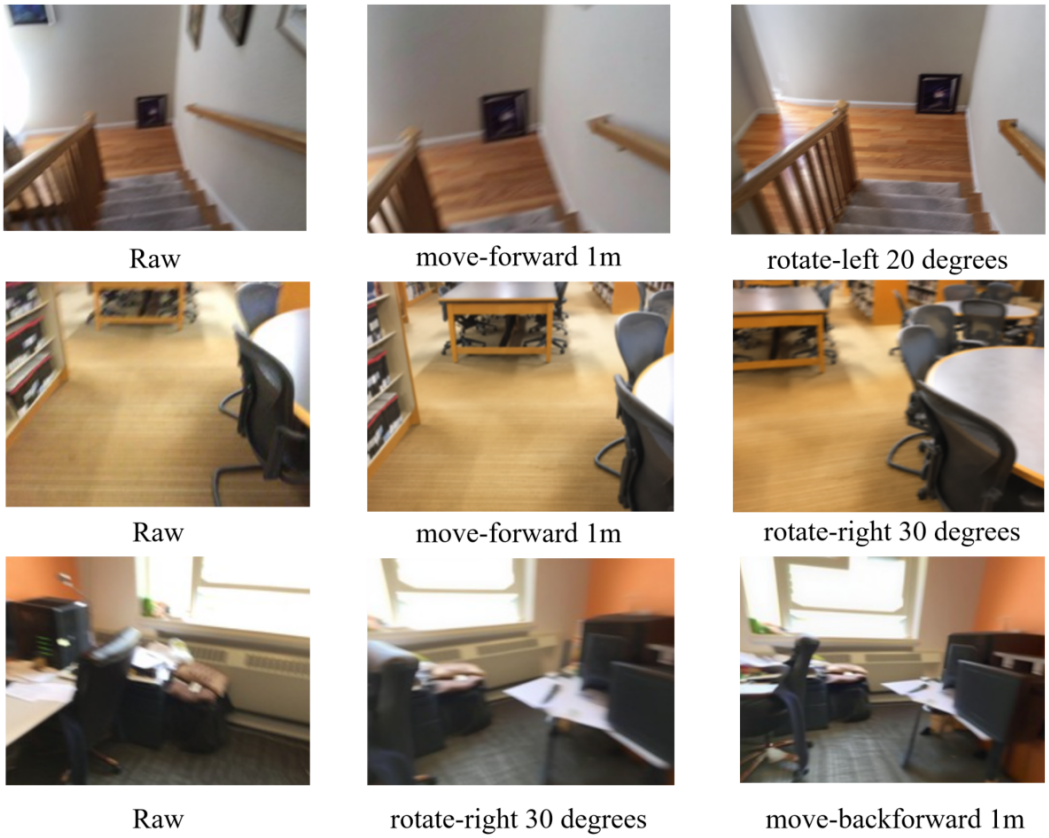


Figure 9: Novel-view observations synthesized through action-conditioned multi-hop exploration.

Q:How many laptop(s) appear?

A:<think>I think through the visible views first to locate all potential laptops in the room and avoid double-counting the same object across different viewpoints. In view7, I identify one open laptop placed on the desk near the glass doors. Its screen-and-keyboard structure and its position on the desk surface make it a clear laptop candidate. In view8, I identify another laptop-like object resting on the ottoman next to the sofa. This second object appears closed and is located in a different part of the room, so it should not be merged with the laptop seen on the desk.
I use <segment_start>
<segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_end> to confirm the object boundary of the open laptop on the desk surface, and <segment_start>
<segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_end> to isolate the closed laptop on the ottoman. These segmentation cues help verify that the two laptop candidates correspond to two distinct object regions rather than repeated observations of the same item.
I then verify their spatial separation using
<depth_start><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_end>
and
<depth_start><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_end>.
The depth cues show that the desk-side laptop and the ottoman-side laptop occupy different locations relative to the surrounding furniture. Since one laptop is on the desk near the glass doors and the other is on the ottoman beside the sofa, they are separate entities in the room. Therefore, the total number of laptops is two.</think> <answer>2</answer>

GT:2

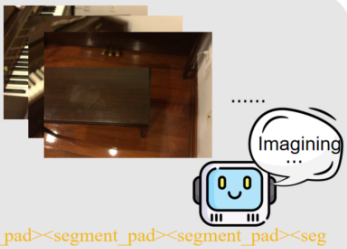


Figure 10: Case study: object counting under partial occlusion. The model invokes segmentation and depth facets to disambiguate overlapping instances.

Q:I am standing in front of the TV putting on a song to brighten my dayWhat is the first object to my immediate right?.

A:<think>I think through the 16 RGB views to establish the room layout, identifying a central table surrounded by chairs and a cabinet against one wall. I use
<depth_start><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_end> to settle the front-back order between the cabinet and the adjacent chair, confirming the chair is positioned immediately next to the cabinet. I then reason from <segment_start>
<segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_end> to keep the spatial relationship fixed across the scene, showing the chair is the closest object to the cabinet's side. I trace the orientation with <normal_start>
<normal_pad><normal_pad><normal_pad><normal_pad><normal_pad><normal_pad><normal_pad><normal_pad><normal_pad> <normal_end>
so the spatial picture stays consistent, verifying the chair is the immediate obstacle when moving away from the cabinet. I bring in
<bev_start><bev_pad><bev_pad><bev_pad><bev_pad><bev_pad><bev_pad><bev_pad><bev_pad><bev_end> to confirm the global layout where the chair sits directly beside the cabinet along the wall. I narrow this down by observing that walking straight back from the cabinet leads directly into the chair.</think> <answer>chair</answer>

GT:chair

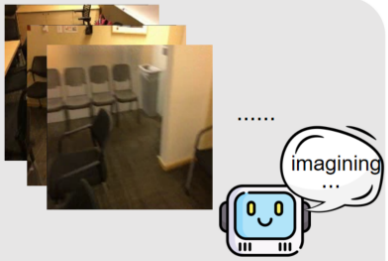
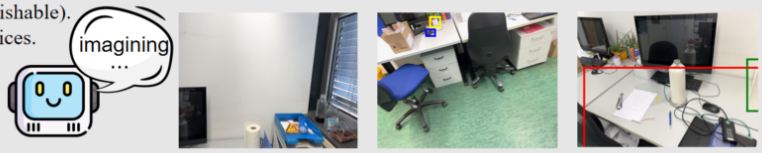


Figure 11: Case study: egocentric navigation. The model relies on BEV and novel-view facets to plan a route from the agent's situated viewpoint.

Q:What is the observer's perspective of table (red bbox) and pen holder (yellow bbox) before and after observer relocating to paper towel (green bbox) and facing mouse (blue bbox)? Calculate or judge based on the 3D center points of these objects. Respond with the observer's perspective in mind, identifying the first image as the main view before any action is taken. Please use the world coordinate system to determine the up and down position relationship of objects.
For multiple-choice questions, consider only the state after the observer has moved.
The options describe the spatial relationship between two objects in terms of left-right (left, right, or empty if indistinguishable), above-below (above, below, or empty if indistinguishable), closer-farther (closer, farther, or empty if indistinguishable), the first object is in front of or behind the observer (front, behind, or empty if indistinguishable), the second object is in front of or behind the observer (front, behind, or empty if indistinguishable).
Pick the right response from the available choices.
A. right, below, closer, behind, front
B. right, , closer, behind, behind
C. right, below, closer, front, front
D. left, below, , front, behind
Your answer can only include one of options A, B, C or D.?



A:<think>I think through the source view first and use <depth_start><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_pad><depth_end> to settle the initial front-back order, identifying the table surface as the primary reference plane for depth comparison. I then reason from <segment_start> <segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_pad><segment_end> to keep the referenced objects fixed, noting that the pen holder sits on top of the table while the mouse rests on the desk surface further back. I bring in <mv_start><mv_pad><mv_pad><mv_pad><mv_pad><mv_pad><mv_pad><mv_pad><mv_pad><mv_end> to simulate the observer relocating to the paper towel area and facing the mouse; this novel view confirms that from this new vantage point, the table structure containing the pen holder appears to the right of the mouse's position. The vertical analysis shows the pen holder is physically higher than the mouse due to its placement on the table edge, establishing a "below" relationship for the mouse relative to the pen holder. The depth cues indicate the table edge is closer to the viewer than the far wall where the mouse is located, making the table "closer". Finally, the spatial arrangement places the pen holder in front of the observer's line of sight towards the mouse when considering the specific orientation of the desk setup.</think> <answer>C</answer>

GT:C

Figure 12: Case study: viewpoint-dependent spatial comparison. The model activates novel-view and depth facets to verify the relative order of objects under occlusion.

Question-Aware Spatial Keyframe Selection

Role: Q&A-aware keyframe selector for spatial reasoning data construction.
Input: A question Q, a ground-truth answer A, and N sampled video frames indexed from 1 to N.
Task: Select a compact set of keyframes that best supports the question-answer pair and preserves useful spatial evidence for constructing auxiliary spatial representations.
Selection criteria:
1. Select frames that directly support the question-answer pair.
2. Prefer frames with spatial evidence: object location, count, depth order, distance, size, occlusion, region boundaries, layout, direction, route structure, or viewpoint changes.
3. Avoid redundant, blurry, transitional, or weakly relevant frames.
4. Select the minimal complementary subset and do not infer invisible evidence.
Output: JSON only. Frame indices must be integers between 1 and N.
`{"selected_frame_indices": [1, 3, 5], "selected_keyframes": [...]}`

Figure 13: Prompt used for question-aware spatial keyframe selection.

Spatial-Tagged Reasoning Trajectory Filtering

Role: Quality filter for spatial-tagged reasoning trajectories.
Input: A question Q, its ground-truth answer, the original RGB observation, and a generated trajectory.
Task: Decide whether to keep or delete the trajectory.
Criteria:
1. The trajectory must directly address Q.
2. The trajectory must support the ground-truth answer.
3. The trajectory must be grounded in the RGB observation.
4. Spatial evidence must be meaningful (location, distance, size, occlusion, boundaries, direction, layout, route, viewpoint).
5. No hallucinated objects, relations, layouts, viewpoints, or spatial details.
6. Spatial tags must match the evidence type: `<depth_start>`, `<normal_start>`, `<segment_start>`, `<bird_start>`, `<mv_start>`.
7. Reject weak grounding, wrong tags, contradictions, irrelevant content, wrong answers, or format errors.
Output: `keep` or `delete`.

Figure 14: Prompt used for filtering low-quality spatial-tagged reasoning trajectories.

Initial Camera Movement Planning

Role: Expert in 3D vision and spatial reasoning.
Input: An image, a spatial question Q, and its answer A.
Task: Plan exactly one camera movement to generate a novel viewpoint that helps answer Q.
Action space: `rotate-{left,right,up,down}` (in-place rotation, parameter: `degree`); `move-{left,right,forward,backward}` (pure translation, parameter: `distance` in meters).
Output format: JSON object: `{"action": ..., "degree": 0, "distance": 1.0, "description": "..."}`. Output only the JSON object.

Figure 15: Prompt used for planning the initial camera movement.

Multi-Step Camera Movement Decision

Role: Expert in 3D vision and spatial reasoning.
Input: Multiple images of the same scene from different viewpoints in chronological order, the question Q, and its answer A.
Task: Decide whether to plan another camera movement or commit to the final answer.
Strategy: Typically 2–3 movements are needed for a complete spatial understanding. Continue exploring if relationships are uncertain; only commit when highly certain or when the step budget is exhausted.
Action space: Same as Figure 15.
Output: JSON, either `{"action": ..., "degree": ..., "distance": ..., "description": "..."}` for continued exploration, or `{"final_answer": "A/B/C/D", "reasoning": "..."}` to commit.

Figure 16: Prompt used for multi-step camera exploration decisions.

Novel-View Image Generation

Role: Novel-view image generation model.
Input: A current-view image and a camera movement instruction.
Task: Generate a new image from the updated viewpoint.
Constraints: Preserve scene identity, object shapes, materials, room structure, furniture layout, lighting, and photographic style. Do not introduce or remove key objects, and avoid severe deformation or physical intersection.
Output: A single image from the updated viewpoint.

Figure 17: Prompt used for generating novel-view images.

Spatial-Tagged Reasoning Trajectory Synthesis

Role: Multi-view spatial reasoning chain generator.
Input: A question Q, the correct answer, and the auxiliary evidence set $\mathcal{F}$ comprising RGB views, depth maps, segmentation maps, normal maps, a BEV layout, and novel-view images.
Reasoning requirements:
1. Integrate RGB views to understand main objects, layout, and viewing direction.
2. Use depth evidence for distance, near/far order, and front/back relations.
3. Use segmentation for object regions, boundaries, and occlusion.
4. Use normal evidence for surface orientation and local geometry.
5. Use BEV evidence for global layout, top-down relations, and route structure.
6. Use novel-view evidence for occlusion verification and viewpoint-dependent relations.
7. Ground reasoning in visual evidence; do not hallucinate.
8. Support the given correct answer; do not mention wrong options; avoid hedging.

Tag rules: Use `<depth_start>viewN<depth_end>`, `<segment_start>viewN<segment_end>`, `<normal_start>viewN<normal_end>`, `<bird_start>bird_1<bird_end>`, `<mv_start>novelN<mv_end>`. Tags appear only inside `<think>...</think>`.
Output: `<think>...</think><answer>...</answer>`, no extra text.

Figure 18: Prompt used for synthesizing spatial-tagged reasoning trajectories with multi-facet evidence.