GS-MLLM: COGNITIVELY INSPIRED TRIADIC PERCEPTION FOR MULTIVIEW SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in spatial MLLMs increasingly leverages geometric priors from 3D vision models. However, exposing reconstructed geometry to an MLLM does not resolve a fundamental problem: *how should geometric evidence be organized for reasoning?* Geometry must associate repeated observations across views without erasing view-specific details, while a limited reasoning capacity must preserve both question-relevant evidence and global scene structure. We introduce GS-MLLM, a triadic perception framework for multimodal spatial reasoning that organizes reconstructed geometry into reliability-aware physical groups and combines question-conditioned local evidence with global scene context. These groups consolidate geometrically consistent observations across views while retaining their source-view evidence. A question-conditioned semantic tokens retrieves task-relevant local detail, while a question-independent adaptive spatial prefix summarizes the global scene layout. Position-aware cross-attention integrates the retrieved geometry into Pixel-token states while preserving the MLLM's pretrained visual-token structure. With Qwen3-VL-4B, GS-MLLM achieves an overall score of **54.7** on VSI-Bench, a **6.0%** improvement over the previous best result, and delivers a **5.4%** improvement in average score over GeoThinker across four additional spatial reasoning benchmarks. These results suggest that effective multimodal spatial reasoning benefits from coordinating visual appearance, task semantics, and scene geometry through a shared spatial reference.

1 INTRODUCTION

Spatial reasoning requires multimodal large language models (MLLMs) to relate partial observations across views. Estimating distances, comparing object positions, and interpreting scene layout are essential for embodied assistance and navigation (Hong et al., 2023; Huang et al., 2023). Current MLLMs excel at semantic understanding (Li et al., 2024; Bai et al., 2025a), yet remain below human performance on spatial reasoning, where linguistic strategies alone offer limited gains (Yang et al., 2025a). This gap prompts us to ask: *How do humans maintain a coherent understanding of a scene across changing viewpoints?* Studies of human navigation provide evidence for spatial mapping (Doeller et al., 2010). Taking this as a design motivation, we seek to retain what each view shows, identify what the question requires, and place the evidence in a common scene reference.

However, existing methods address individual parts: depth augments local geometry (Cheng et al., 2024), reconstruction models recover dense scene structure (Zheng et al., 2026), and geometry-aware decoders selectively access geometric evidence (Li et al., 2026b). Yet they leave open how these signals should be organized for multiview reasoning. Two tensions are central: cross-view consolidation can obscure source-view evidence, whereas keeping views separate fragments spatial relations; task-conditioned selection can improve relevance, but discarding context reduces global scene coverage. The design challenge is therefore to organize cross-view geometry while preserving task-relevant detail and the global context needed to interpret it (Fig. 1).

Motivated by spatial mapping and functional accounts of visual processing and semantic memory (Mishkin et al., 1983; Binder & Desai, 2011), we formulate *triadic perception* through three interacting roles. Pixel representations preserve the visual appearance of each observation, Semantic representations express the task and guide attention toward relevant evidence, and Spatial representations provide a shared reference for relating observations across views. These roles are

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.

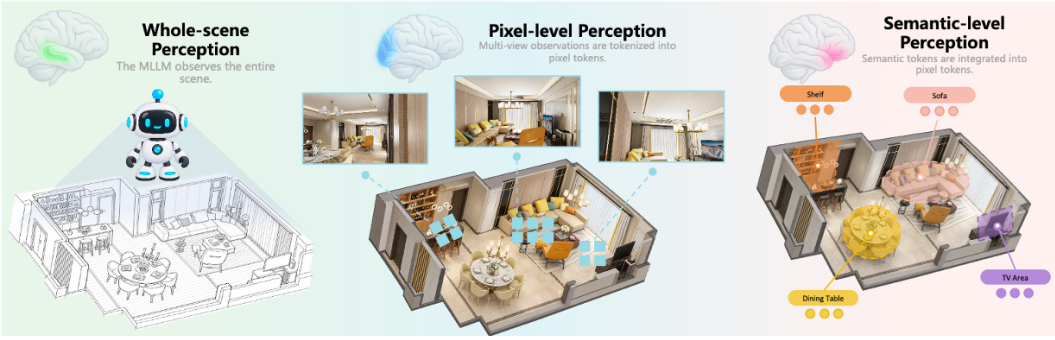


Figure 1: **Triadic perception for multiview spatial reasoning.** The three functional roles are complementary: spatial representations organize observations from multiple views into a shared scene reference, Pixel representations preserve view-specific visual appearance, and Semantic representations connect the scene to task-relevant concepts and regions.

realized through three operations: organizing reconstructed geometry into reliability-aware *physical groups*; accessing it through complementary local task-conditioned retrieval and a global task-independent *adaptive spatial prefix*; and fusing retrieved local geometry back into Pixel-token states with coordinate-grounded, view-aligned attention. This interface keeps question-relevant detail and global scene context available within a shared spatial reference. With Qwen3-VL-4B, GS-MLLM achieves **54.7** on VSI-Bench, exceeding GeoThinker's corresponding variant by **3.1 points** (Table 1). Across MMSI-Bench, ViewSpatial, CV-Bench, and MindCube, it improves over the same baseline on all four benchmarks, raising the average score from 47.9 to **50.5**, a **2.6-point** gain (Table 2). Component ablations further support the roles of physical grouping, question-guided retrieval, global spatial context, and coordinate-grounded fusion (Sec. 4.4).

Our contributions are threefold:

- **Cognitively inspired triadic perception.** We introduce GS-MLLM, a scene-centric framework that coordinates Pixel, Semantic, and Spatial representations, connecting view-specific appearance and task relevance through a shared spatial reference.
- **Scene-organized local–global access.** We organize cross-view geometry with reliability-aware physical groups and expose it through complementary question-guided retrieval and an adaptive spatial prefix, followed by coordinate-grounded, view-aligned fusion.
- **Empirical validation across tasks and backbones.** Evaluation on five spatial benchmarks with two MLLM backbones demonstrates aggregate gains over GeoThinker, while component ablations support the roles of scene organization, selective access, and geometric grounding.

2 Related Work

Multimodal Models and Spatial Supervision. Modern MLLMs couple pretrained vision encoders with language models (Alayrac et al., 2022; Li et al., 2023) and improve their visual–language interface through instruction tuning and alignment (Dai et al., 2023; Liu et al., 2023; Zhu et al., 2024). Models such as InternVL (Chen et al., 2024c), LLaVA-OneVision (Li et al., 2024), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a) extend this capability to multi-image and video inputs. Spatial reasoning further requires observations from different views to be interpreted as parts of one physical scene. SpatialVLM (Chen et al., 2024a), SPAR (Zhang et al., 2026b), SpatialLadder (Li et al., 2026c), and Cambrian-S (Yang et al., 2026a) improve this ability through spatial supervision or specialized training. These methods primarily change the data or objective used to learn spatial behavior. Our work complements this line by focusing on how reconstructed geometry is organized and exposed to the MLLM, rather than introducing another spatial training recipe.

Geometry-Aware Spatial Reasoning. A complementary line of work supplies explicit geometric priors. SpatialRGPT (Cheng et al., 2024) adds depth to region features, while LLaVA-3D (Zhu et al., 2025a) adds 3D positions to visual patches. VGGT (Wang et al., 2025) provides dense geometry

from RGB views; VG-LLM (Zheng et al., 2026) and Spatial-MLLM (Wu et al., 2026) incorporate such features with different encoder designs. Other methods use video or camera geometry (Fan et al., 2026; Zhang et al., 2026a), jointly reconstruct and reason in 3D (Hu et al., 2026), or impose cross-view consistency (Yeh et al., 2026). These studies establish the value of geometric priors but leave open how repeated observations should be organized across views while retaining view-specific evidence. GS-MLLM addresses this representation question with reliability-aware physical groups. We therefore treat cross-view correspondence as part of the model interface, using physical groups as the unit on which relevance is evaluated.

Selective Access to Visual and Geometric Evidence. Recent work also controls which evidence reaches the language model. FastV (Chen et al., 2024b), SparseVLM (Zhang et al., 2024), and VisionZip (Yang et al., 2025b) select or compress visual tokens, while GSM (Zhou et al., 2026) and GeoThinker (Li et al., 2026b) extend selective access to geometric features. GeoThinker in particular enables semantic states to retrieve geometry through frame-strict interaction. GS-MLLM differs in where selection occurs: question tokens score physical groups rather than independent patches, after which local source observations are retrieved and a question-independent spatial prefix summarizes the scene layout. This combines task-directed detail with global spatial context. It also separates the choice of what to inspect from the scene summary needed to interpret the selected evidence.

3 METHOD

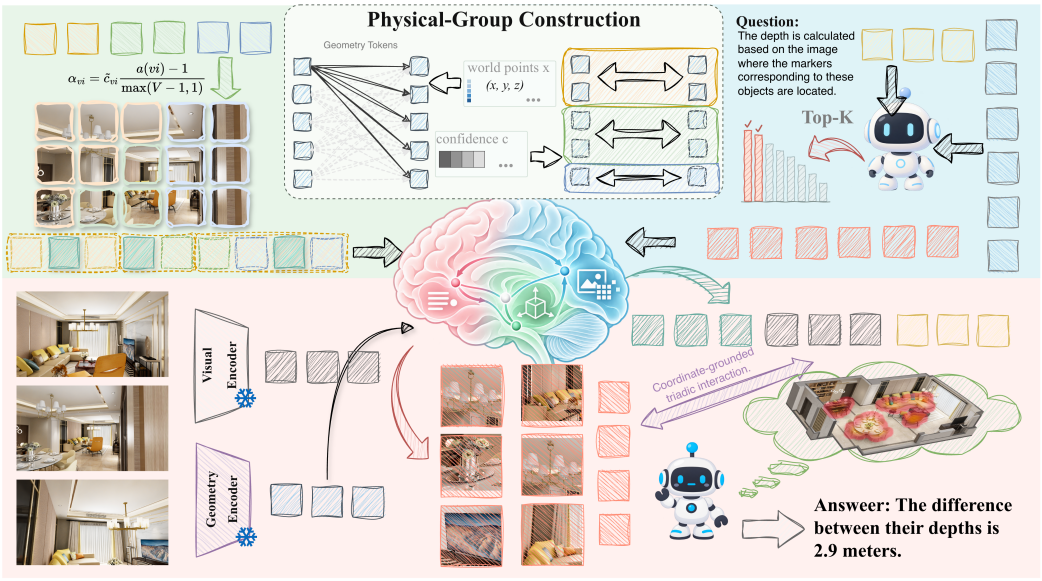


Figure 2: **Overview of GS-MLLM.** Frozen visual and geometry encoders extract appearance and geometric features from the input views. Reliability-aware physical groups consolidate corresponding geometric observations across views. Question-guided top-K retrieval selects task-relevant local geometry, while a question-independent spatial prefix summarizes the global scene layout. Coordinate-grounded, view-aligned attention integrates the selected geometry into Pixel-token states, and the spatial prefix supplies global context for multimodal reasoning.

3.1 OVERVIEW: TRIADIC PERCEPTION

Given V views $\mathcal{I} = I_{v}{}_{v=1}^{V}$ and a question Q, our method organizes perception into three functional roles: *Pixel tokens* preserve appearance on the pretrained visual grid, *Semantic tokens* capture the question and evolving reasoning states, and *Spatial tokens* encode reconstructed scene geometry.

We expose geometry at two complementary levels. Geometric observations are first consolidated into reliability-aware physical groups that form a *hard geometry memory*. Semantic tokens selec-

tively retrieve task-relevant geometry from this memory, while a question-independent *adaptive spatial prefix* provides a compact summary of the global scene layout. Retrieved geometry is then integrated into Pixel-token states through position-aware attention. Together, these components preserve the pretrained Pixel–Semantic pathway while introducing structured spatial evidence for downstream reasoning.

3.2 Reliability-Aware Spatial Token Construction

Dense geometric observations. For each view, a frozen geometry encoder yields patch features $\mathbf{G}_v = \{\mathbf{g}_{vi}\}_{i=1}^{P} \in \mathbb{R}^{P\times d_g}$. We confidence-pool its dense world-point predictions onto the patch grid to obtain a point $\mathbf{x}_{vi} \in \mathbb{R}^3$ and confidence c_{vi} per patch. The predicted camera center determines the unit ray $\mathbf{r}_{vi}$. All views share a scene normalization. Let $\boldsymbol{\mu}$ be the coordinate-wise median of valid points and s the 90th percentile of their distances to $\boldsymbol{\mu}$. We center points at $\boldsymbol{\mu}$ and divide by $\max(s, \epsilon)$ to obtain $\widetilde{\mathbf{x}}_{vi}$. Confidences are divided by $\max(q_{0.9}(c), \epsilon)$ and clipped to $[0, 1]$ to obtain $\widetilde{c}_{vi}$, where $q_{0.9}$ denotes the 90th percentile. This makes spatial matching invariant to scene translation and scale.

Scene-level physical grouping. We associate observations of the same surface across views into physical groups while retaining their source patches. Geometry features are mapped into the selector space as $\mathbf{z}_{vi} = \mathrm{norm}(\mathbf{W}_g\,\mathrm{LN}(\mathbf{g}_{vi}))$. For every pair of views, candidate correspondences are reciprocal nearest neighbors in normalized 3D space. We retain a pair (vi, uj) only if its distance satisfies $\|\widetilde{\mathbf{x}}_{vi} - \widetilde{\mathbf{x}}_{uj}\|_2 \le \delta_x$ and its feature similarity satisfies $\mathbf{z}_{vi}^{\top}\mathbf{z}_{uj} \ge \delta_f$. Retained pairs are merged by constrained union–find, allowing at most one observation per view in each group. This yields the patch-to-group assignment $a(vi) \in \{1, \dots, G\}$.

Group r has canonical feature $\overline{\mathbf{g}}_r$ and centroid $\overline{\mathbf{x}}_r$, computed by averaging $\mathbf{g}_{vi}$ and $\widetilde{\mathbf{x}}_{vi}$ over its members with weights $\widetilde{c}_{vi}$. Its reliability is $C_r = 1 - \prod_{vi:a(vi)=r}(1 - \widetilde{c}_{vi})$.

Each local feature is updated with confidence-weighted group consensus:

$$\widetilde{\mathbf{g}}_{vi} = \mathbf{g}_{vi} + \alpha_{vi}(\overline{\mathbf{g}}_{a(vi)} - \mathbf{g}_{vi}), \qquad \alpha_{vi} = \widetilde{c}_{vi}\frac{n_{a(vi)} - 1}{\max(V - 1, 1)}, \tag{1}$$

where n_r counts the distinct supporting views. Singletons retain their original features; reliable multi-view matches receive stronger consensus.

3.3 Semantic–Spatial Geometry Memory

Question-aware hard retrieval. We layer-normalize and project the latest question's embeddings into the same selector space, yielding normalized Semantic vectors $\mathbf{q}_\ell$. Their affinity with canonical group features $\overline{\mathbf{z}}_r$ is $A_{r\ell} = \overline{\mathbf{z}}_r^{\top}\mathbf{q}_\ell/\tau$. Following SparseVLM's text-rater intuition (Zhang et al., 2024), we select informative question tokens by their mean attention from groups:

$$e_\ell = \frac{1}{G}\sum_{r=1}^{G}\mathrm{softmax}_\ell(A_{r\ell}), \qquad m_\ell = \mathbf{1}[e_\ell \ge \mathrm{mean}(\mathbf{e})]. \tag{2}$$

Here, $\mathbf{1}[\cdot]$ equals one when the condition holds and zero otherwise. Selected raters receive uniform forward weights w_ℓ, with $\mathrm{softmax}(\mathbf{e}/\tau_r)$ supplying straight-through gradients. They vote over groups through reverse text-to-geometry attention:

$$p_r = \sum_{\ell} w_\ell\,\mathrm{softmax}_r(A_{r\ell}), \qquad s_r = \log(Gp_r) + \beta\log(1 + n_r). \tag{3}$$

Here, $\log(Gp_r)$ measures relevance relative to a uniform group distribution, and the consensus term favors groups supported by multiple views.

Every patch inherits its group's relevance, with final priority $u_{vi} = s_{a(vi)} + \gamma\log(\max(\widetilde{c}_{vi}, 10^{-4}))$. We select the top K valid patches independently per view, with K equal to that view's final Pixel-token count after spatial merging and optional long-sequence downsampling. The selected values are the local features from Eq. 1. A straight-through top-K gate uses hard forward selection and softmax backward gradients, training the selector solely through the language-modeling objective.

Global spatial prefix. We summarize the observed 3D layout using all valid physical groups, independently of the question. A pointwise LayerScale residual MLP produces refined features $\widehat{\mathbf{g}}_r$ using each group's position, reliability, view support, observation count, and feature dispersion (Appendix C.2). For G valid groups, the prefix contains $K_p = \max(1, \lceil \rho G \rceil)$ anchors. We choose the first center by maximum confidence and subsequent centers by confidence-weighted farthest-point sampling. Every group is assigned to its nearest center in normalized 3D space, and the features, centroids, and confidences in each region are confidence-pooled. Each anchor sums the linear feature projection $\mathbf{W}_a \widehat{\mathbf{g}}_k$, the position–confidence embedding $\phi([\overline{\mathbf{x}}_k, C_k])$ from a two-layer MLP, and a type embedding $\mathbf{e}_{\text{type}}$, all at the MLLM width. We insert the anchors immediately before the first vision-start token and mask their language-modeling loss labels. This exposes scene context to subsequent causal layers while preserving the delimited Pixel-token layout.

3.4 Triadic Geometry Integration

Local 3D memory consolidation. Selected Spatial tokens gather context by attending to raw patch features from the geometry encoder within their source view. Both are projected to the MLLM width, and a shared MLP embeds normalized 3D coordinates into queries and keys:

$$\mathbf{H}'_v = \mathbf{H}_v + \tanh(\eta)\,\mathrm{MHA}\big(\mathbf{H}_v + \lambda_x \phi_x(\mathbf{X}_v^H), \mathbf{P}_v + \lambda_x \phi_x(\mathbf{X}_v^P), \mathbf{P}_v\big)\,. \tag{4}$$

Here, $\mathbf{H}_v$ and $\mathbf{P}_v$ denote selected and raw patch features, respectively, and η is a learnable gate. Invalid points are masked. A bottleneck MLP with a small LayerScale residual follows the attention update.

Coordinate-grounded triadic interaction. We embed scene-normalized dense (x, y, z, c) maps using a convolution whose kernel and stride equal the vision patch size times the spatial-merge factor. The resulting $\mathbf{E}_v^{xyz}$ aligns one-to-one with the final Pixel-token grid and is added only to the Pixel-token query source. Using these coordinate-grounded queries, we follow GeoThinker's active geometry integration (Li et al., 2026b) and inject geometry after self-attention and before the feed-forward block in the middle half of the language decoder. At layer ℓ, only states $\mathbf{S}_v^\ell$ at visual-token positions are updated through a learnable gate g_ℓ:

$$\mathbf{S}_v^\ell \leftarrow \mathbf{S}_v^\ell + \tanh(g_\ell)\,\mathrm{Attn}\big(\mathbf{S}_v^\ell + \mathbf{E}_v^{xyz}, \mathbf{H}'_v, \mathbf{H}'_v\big)\,. \tag{5}$$

Queries attend only to geometry from the same view; cross-view consensus is already encoded through physical groups (Sec. 3.2).

4 Experiments

4.1 Experimental Setup

Data and benchmarks. We follow the training regime used by prior video-based geometry models: 234K spatial instructions from SPAR-7M (Zhang et al., 2026b) and 64K general video instructions from the LLaVA-Hound (Zhang et al., 2025) subset released with VG-LLM (Zheng et al., 2026). VSI-Bench (Yang et al., 2025a) is our primary benchmark: its overall score averages mean relative accuracy on four numerical tasks and exact-match accuracy on four multiple-choice task groups. We also evaluate MMSI-Bench (Yang et al., 2026b), ViewSpatial (Li et al., 2026a), CV-Bench (Tong et al., 2024), and MindCube (Yin et al., 2025), their reported average weights the four benchmarks equally. SPBench (Li et al., 2026c) provides a separate analysis by input setting. Benchmark definitions and aggregation rules appear in Appendix B.1.

Baselines. We compare with proprietary MLLMs, open-source generalists, and spatially specialized models. GeoThinker (Li et al., 2026b) is the closest geometry-enhanced baseline; we emphasize comparisons with its corresponding Qwen2.5-VL-3B (Bai et al., 2025b) and Qwen3-4B (Bai et al., 2025a) variants. These comparisons evaluate full models; comparisons with an untuned base MLLM additionally include the effect of spatial instruction tuning.

Model and training. GS-MLLM uses a frozen VGGT-1B geometry encoder (Wang et al., 2025) and a frozen semantic vision tower; the language model and geometry integration modules are optimized for one epoch. Results on VSI-Bench and the four additional benchmarks use Qwen2.5-VL-3B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a). Video evaluation uniformly samples

Table 1: **Performance on VSI-Bench under the training regime.** Results are grouped by model family. Best and second-best populated averages are shown in **bold** and underlined, respectively.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baseline									
Chance Level (Random)	–	–	–	–	–	25.0	36.1	28.3	25.0
Chance Level (Frequency)	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Proprietary Models (API)									
GPT-4o (Hurst et al., 2024)	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash (Team et al., 2024)	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro (Team et al., 2024)	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-Source Models									
LLaVA-OneVision-7B (Li et al., 2024)	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B (Li et al., 2024)	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B (Liu et al., 2024)	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B (Liu et al., 2024)	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
InternVL2-8B (Chen et al., 2024c)	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B (Chen et al., 2024c)	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
Qwen2.5VL-3B (Bai et al., 2025b)	28.6	32.7	19.5	17.3	25.1	37.3	44.9	30.4	21.8
Qwen2.5VL-7B (Bai et al., 2025b)	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
Open-Source Spatial Intelligence Models									
SPAR-8B (Zhang et al., 2026b)	44.1	–	–	–	–	–	–	–	–
SpatialLadder-3B (Li et al., 2026c)	44.8	–	–	–	–	–	–	–	–
Spatial-MLLM-4B (Wu et al., 2026)	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VG-LLM-4B (Zheng et al., 2026)	46.7	67.6	37.6	55.2	52.5	48.0	44.7	31.9	35.5
VG-LLM-8B (Zheng et al., 2026)	49.7	68.1	38.7	59.0	61.1	45.5	44.9	26.8	53.4
CA-MLLM-4B (Zhang et al., 2026a)	46.8	71.3	39.3	50.4	65.9	50.7	45.5	31.4	20.2
GeoThinker$_{\text{Qwen2.5VL-3B}}$ (Li et al., 2026b)	47.8	67.5	37.0	56.9	62.3	45.0	47.5	32.4	33.9
GeoThinker$_{\text{Qwen2.5VL-7B}}$ (Li et al., 2026b)	50.5	69.5	38.5	57.9	62.2	45.2	46.2	31.4	52.6
GeoThinker$_{\text{Qwen3-4B}}$ (Li et al., 2026b)	<u>51.6</u>	65.7	42.0	63.0	59.8	53.2	47.0	30.9	51.5
GS-MLLM$_{\text{Qwen2.5VL-3B}}$	49.2	68.1	36.2	57.9	60.9	44.9	47.0	35.1	43.2
GS-MLLM$_{\text{Qwen3-VL-4B}}$	**54.7**	70.7	42.7	65.1	65.9	54.8	49.2	29.4	59.4

"–" denotes an unavailable or unreported entry. Obj., Abs., Rel., and Appr. abbreviate object, absolute, relative, and appearance, respectively.

up to 32 frames with greedy decoding. Backbone configurations, training frame counts, and optimization settings are detailed in Appendix B.2.

4.2 Main Results on VSI-Bench

The aggregate gain holds across both backbones. Table 1 shows that GS-MLLM reaches **54.7** with Qwen3-VL-4B, the highest average among the compared models and **+3.1 points** above GeoThinker's Qwen3-4B variant (51.6). The same direction holds with Qwen2.5-VL-3B: GS-MLLM scores **49.2**, exceeding GeoThinker with that backbone by **+1.4 points**. The 4B result also exceeds the reported scores of GeoThinker with Qwen2.5-VL-7B (50.5) and VG-LLM-8B (49.7). **Appearance order shows the largest gain across both backbones.** Relative to the corresponding GeoThinker variants, GS-MLLM achieves its largest task-level gain in appearance order on both backbones: +9.3 points with Qwen2.5-VL-3B and +7.9 with Qwen3-VL-4B. The 4B model improves on seven of eight task groups, including all four numerical tasks, with further gains in room size (+6.1) and object counting (+5.0). The 3B model improves on four of eight task groups; route planning improves by 2.7 points with 3B but declines by 1.5 with 4B.

The advantage extends beyond VSI-Bench. GS-MLLM achieves four-benchmark averages of **50.50** with Qwen3-VL-4B and **49.28** with Qwen2.5-VL-3B (Table 2). These exceed the corresponding GeoThinker variants by **+2.60 points** and **+3.50 points** and are the two highest averages among the listed open-source models. Both remain below the listed proprietary models.

Mental modeling shows the largest recurring gain. The Qwen3-VL-4B model improves over GeoThinker on all four benchmarks; the Qwen2.5-VL-3B model improves on three. In both cases, the largest gain is on MindCube: +5.1 points with Qwen3-VL-4B and +10.4 with Qwen2.5-VL-3B. MindCube evaluates cognitive mapping, perspective-taking, and mental simulation from partial views; improvement on both backbones is consistent with the goal of a shared scene representation. The benefit still depends on the backbone and task: ViewSpatial improves by 1.9 points with Qwen3-VL-4B but declines by 1.8 with Qwen2.5-VL-3B. On CV-Bench, which uses single-image inputs, GS-MLLM with Qwen3-VL-4B scores 83.6, above GeoThinker (81.5) but below its base MLLM (85.6). This variation motivates the following comparison of single-image spatial judgments and spatial scene understanding.

Table 2: Overall accuracy (%) on four spatial reasoning benchmarks. Avg. is the unweighted mean over all four benchmarks and is reported only when all four scores are available. In the Avg. column, the best and second-best results among open-source models are shown in **bold** and underlined, respectively.

METHODS	AVG.	MMSI-BENCH	VIEWSPATIAL	CV-BENCH	MINDCUBE
BASELINES					
Human	–	97.2	–	–	94.5
Random Choice	–	25.0	26.3	–	33.0
PROPRIETARY MODELS					
Seed-1.6 (Guo et al., 2025)	54.00	38.3	43.8	85.2	48.7
Gemini-2.5-Pro (Comanici et al., 2025)	56.88	38.0	46.0	85.9	57.6
GPT-5 (Singh et al., 2025)	57.05	41.8	45.5	84.6	56.3
Gemini-3-Pro-Preview (Team et al., 2023)	64.58	45.2	50.3	92.0	70.8
OPEN-SOURCE GENERALIST MODELS					
BAGEL-7B-MoT (Deng et al., 2025)	45.75	31.0	41.3	76.0	34.7
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	42.48	28.6	31.9	71.8	37.6
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	43.75	26.8	36.8	75.4	36.0
Qwen3-VL-2B-Instruct (Bai et al., 2025a)	38.98	11.9	34.2	78.4	31.4
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	44.15	26.7	37.3	85.6	27.0
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	45.68	28.8	39.0	85.1	29.8
InternVL3-2B (Zhu et al., 2025b)	43.25	26.5	32.5	76.5	37.5
InternVL3-8B (Zhu et al., 2025b)	47.28	28.0	38.6	81.0	41.5
OPEN-SOURCE SPATIAL INTELLIGENCE MODELS					
SpatialLadder-3B (Li et al., 2026c)	46.08	27.4	39.8	73.7	43.4
Cambrian-S-3B (Yang et al., 2026a)	42.98	25.2	39.0	75.2	32.5
Cambrian-S-7B (Yang et al., 2026a)	45.80	25.8	40.9	76.9	39.6
VLM-3R-7B (Fan et al., 2026)	45.05	27.9	40.5	71.8	40.0
VG-LLM-4B (Zheng et al., 2026)	47.25	28.6	42.4	81.1	36.9
CA-MLLM-4B (Zhang et al., 2026a)	45.08	27.2	39.3	74.9	38.9
GeoThinker$_{\text{Qwen2.5-3B}}$ (Li et al., 2026b)	45.78	27.1	43.9	72.5	39.6
GeoThinker$_{\text{Qwen3-4B}}$ (Li et al., 2026b)	47.90	26.4	46.2	81.5	37.5
GS-MLLM$_{\text{Qwen2.5-3B}}$	<u>49.28</u>	27.8	42.1	77.2	50.0
GS-MLLM$_{\text{Qwen3-4B}}$	**50.50**	27.7	48.1	83.6	42.6

4.3 WHERE DOES STRUCTURED GEOMETRY HELP?

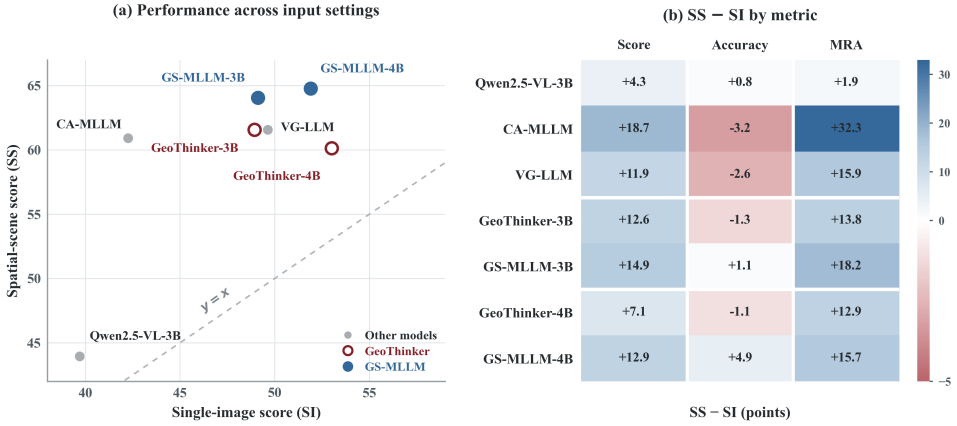


Figure 3: **Spatial scene understanding on SPBench.** (a) Each model's single-image (SI, horizontal axis) and spatial-scene (SS, vertical axis) scores. Subset scores are unweighted means of task-category metrics. Blue markers denote GS-MLLM; red outlines denote GeoThinker. (b) Within-model SS–SI differences in score, multiple-choice accuracy, and numerical MRA, in points. The 3B and 4B GeoThinker and GS-MLLM variants use Qwen2.5-VL-3B and Qwen3-VL-4B, respectively.

We use SPBench to assess spatial scene understanding through its spatial-scene (SS) subset, which evaluates spatial judgments from multiple views of a shared scene. The single-image (SI) subset provides a complementary evaluation of spatial judgments from individual images. **Consistent gains in spatial scene understanding.** Relative to GeoThinker with the same backbone, GS-MLLM gains

0.18 points on SI and **2.49 points** on SS with Qwen2.5-VL-3B (Fig. 3(a)). With Qwen3-VL-4B, it scores **1.12 points lower** on SI and **4.63 points higher** on SS than GeoThinker. These results support a consistent advantage in spatial scene understanding across both backbones, while gains on single-image spatial judgments remain mixed. The metric breakdown in Appendix D.1 reinforces this pattern: relative to GeoThinker, both accuracy and MRA improve on SS for both backbones, whereas SI accuracy falls and MRA rises only slightly. Figure 3(b) summarizes each model's SS–SI differences. Because the subsets contain different questions and task compositions, these gaps provide a descriptive comparison of performance across the two settings.

4.4 Component Ablations

Table 3 examines six components of GS-MLLM with Qwen3-VL-4B on VSI-Bench, covering scene representation, memory access, and geometry integration. The full model scores **54.65** overall and achieves the highest numerical and multiple-choice means among the reported variants.

Local and global memory. The largest overall gap is observed without the adaptive spatial prefix: **2.83 points** below the full model, compared with **1.39 points** without question-aware hard retrieval. The variant without the prefix also has the lowest score in both task families, with a larger gap on multiple-choice tasks (3.96 points) than on numerical tasks (1.69 points). The two branches expose different evidence: the prefix summarizes the global scene layout, while retrieval selects question-relevant local observations.

Table 3: **Component ablations on VSI-Bench.**

Variant	Avg.	Num.	MC
Full GS-MLLM	**54.65**	**61.11**	**48.19**
w/o physical grouping	52.80	59.92	45.68
w/o cross-view consensus	53.19	60.32	46.07
w/o question-aware hard retrieval	53.26	60.40	46.12
w/o adaptive spatial prefix	51.82	59.42	44.23
w/o raw-patch consolidation	54.20	60.84	47.57
w/o dense coordinate queries	53.24	60.31	46.16

Num./MC: means over four numerical/multiple-choice task groups (0–100).

Representation and integration. Variants without physical grouping and cross-view consensus score **1.85** and **1.46 points** below the full model, respectively. These operations associate observations across views and refine their features using shared geometric evidence. For integration, the gap is **1.41 points** without dense coordinate queries and **0.45 points** without raw-patch consolidation. The former grounds Pixel-token queries in 3D, while the latter enriches selected geometry with source-view context.

4.5 Efficiency Analysis

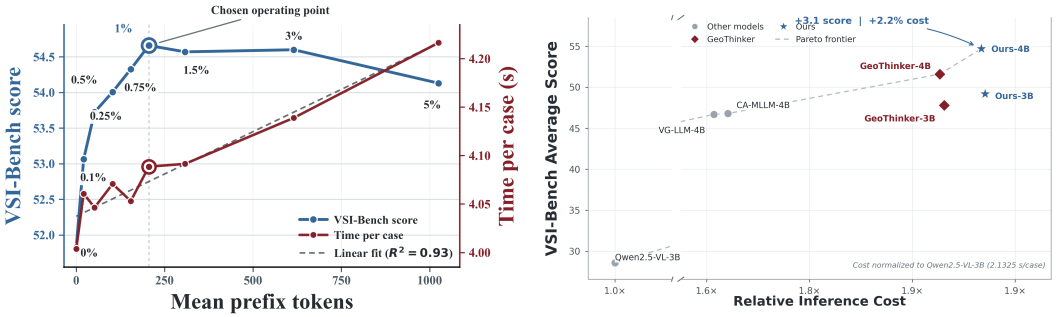
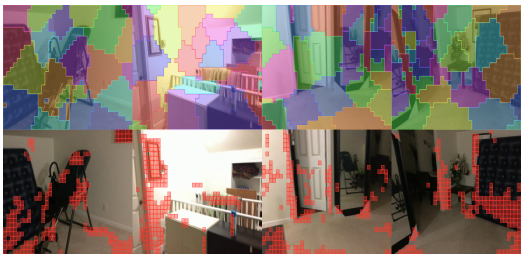


Figure 4: **Performance–cost trade-offs on VSI-Bench. Left:** Score (blue, left axis) and evaluation time per case (red, right axis) versus mean prefix length. Labels show prefix ratios; rings mark the 1% setting. Timings include evaluation overhead; the dashed line shows a linear fit. **Right:** Average score versus reported time per case relative to Qwen2.5-VL-3B (2.1325 s/case). The horizontal axis is broken.

We analyze the performance–cost trade-off on VSI-Bench by varying the spatial-prefix budget and comparing GS-MLLM with other models (Fig. 4). Cost is measured as reported evaluation time per case, including evaluation overhead; the accounting protocol is given in Appendix D.4. **A compact prefix balances performance and cost.** Figure 4 (left) shows that the 1% prefix achieves the highest



Question: If I am standing by the bed and facing the chair, is the mirror to the left or the right of the chair?

Answer: left
Prediction: left



Question: What is the length of the longest dimension of the whiteboard, measured in centimeters?

Answer: 200 cm
Prediction: 200 cm

Figure 5: **Spatial tokens and Semantic-guided selection.** Each case shows four images per row: spatial-prefix areas in distinct colors (top), and patches retrieved by Semantic tokens in red (bottom). Matching areas retain their colors across images within each case.

score among the nine tested budgets, using approximately **205 tokens**. Compared with no prefix, the score reaches **54.66** (**+2.83 points**), while evaluation time is 4.089 s per case (**+2.1%**). Relative to the 1% setting, the 3% budget scores 54.60 (-0.06 points) at 4.139 s per case (+1.2%), and the 5% budget scores 54.13 (-0.53 points) at 4.216 s per case (+3.1%). Across the tested prefix budgets, time grows approximately linearly with prefix length. This pattern supports the 1% budget as a compact operating point: it captures the observed performance gain without the additional cost of larger prefixes. **Higher scores with a small increase in evaluation time.** Figure 4 (right) compares this trade-off across models. Compared with the corresponding GeoThinker variant, GS-MLLM with Qwen3-VL-4B scores **54.7** (**+3.1 points**) at a reported evaluation time of 4.0885 s per case (**+2.2%**). With Qwen2.5-VL-3B, it scores **49.2** (**+1.4 points**) at 4.0963 s per case (**+2.1%**) relative to its GeoThinker counterpart.

4.6 Qualitative Case Studies

Figure 5 shows complementary evidence from spatial-prefix pooling (top) and Semantic-guided retrieval (red, bottom). For relative direction, retrieved patches on the chair, bed rails, and surfaces near the mirror complement the prefix's shared spatial layout; the model correctly locates the mirror to the *left* of the chair when viewed from beside the bed facing the chair. For object size, retrieval covers the whiteboard, its lower boundary, and nearby desk surfaces, combining target detail with context despite changes in apparent size across views. The predicted longest dimension, 200 cm, matches the ground truth. The visualizations highlight two properties. First, views contribute different numbers of prefix tokens according to visible scene coverage, while repeated observations can share 3D anchors (consistent colors), preserving cross-view references. Second, retrieval includes surrounding surfaces and boundaries beyond named objects, providing context for estimating extent and relative position. These overlays do not establish which patches determine the answers. Appendix E provides settings, text-rater selections, and further success and failure cases.

5 Conclusion

We presented GS-MLLM, a scene-centric triadic perception framework that connects visual appearance, task semantics, and reconstructed geometry for multimodal spatial reasoning. Reliability-aware physical groups consolidate geometrically consistent observations across views while retaining source-view details. Building on this shared representation, question-guided retrieval selects task-relevant local evidence, and a question-independent adaptive spatial prefix summarizes the global scene layout. Coordinate-grounded, view-aligned attention integrates the retrieved geometry into Pixel-token states. Experiments suggest that the organization of geometric evidence is an important part of the reasoning interface: associating observations through shared physical regions and exposing complementary local and global information provides a promising direction for multimodal reasoning over multiple views of a scene.

LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used as general-purpose assistance tools during the preparation of this manuscript. Specifically, LLMs were used to assist with drafting and editing portions of the manuscript, improve language clarity and overall readability, and provide assistance with software code, including implementation, debugging, and code refinement. All research ideas, methodological decisions, experimental design, analysis, and experimental results originated from the authors. All AI-assisted text and code were carefully reviewed, verified, and, where necessary, revised by the authors to ensure technical accuracy and consistency with the intended methodology. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide an anonymous repository containing the code and scripts needed to reproduce the reported experiments. The paper and appendices describe the model architecture, reliability-aware physical grouping, question-guided retrieval, spatial-prefix construction, and coordinate-grounded fusion in Sections 3–3.1 and Appendix C. We specify the 234K SPAR-7M spatial-instruction and 64K LLaVA-Hound video-instruction training mixture, benchmark definitions, score aggregation, and evaluation protocols in Appendix B.1; backbone configurations, optimization settings, frame sampling, and inference details are given in Appendix B.2. The reported ablations, sensitivity analyses, efficiency measurements, and qualitative case studies are documented in Appendices D–E.

ETHICS STATEMENT

This work studies multimodal spatial reasoning using publicly available benchmark datasets and publicly released pretrained models. We do not collect new human-subject data or personal information; all experiments use existing data under their original research-use conditions. The proposed method is intended for research on spatial reasoning and is evaluated only in controlled benchmark settings. Although these capabilities may eventually support embodied or assistive applications, incorrect spatial predictions could cause harm; the model should therefore not be used in safety-critical settings without additional validation and safeguards.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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*, 2025a.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jeffrey R Binder and Rutvik H Desai. The neurobiology of semantic memory. *Trends in cognitive sciences*, 15(11):527–536, 2011.

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.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

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.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Christian F Doeller, Caswell Barry, and Neil Burgess. Evidence for grid cells in a human memory network. *Nature*, 463(7281):657–661, 2010.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2 VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9546, 2026.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 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, 2025a.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 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.

Chun-Hsiao Yeh, Shengyi Qian, Manchen Wang, Yi Ma, Joseph Tighe, and Fanyi Xiao. Beyond 3d vqas: Injecting 3d spatial priors into vision-language models for enhanced geometric reasoning. *arXiv preprint arXiv:2605.30231*, 2026.

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

Gongjie Zhang, Wenhao Li, Quanhao Qian, Jiuniu Wang, Deli Zhao, Shijian Lu, and Ran Xu. On the generalization capacities of mllms for spatial intelligence. *arXiv preprint arXiv:2603.06704*, 2026a.

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, 2026b.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

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.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 2026.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305. IEEE, 2025a.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

A CODE

Our code is available at *https://anonymous.4open.science/r/GS-MLLM-FF42*.

B EXPERIMENTAL DETAILS

B.1 BENCHMARKS AND EVALUATION PROTOCOL

VSI-Bench. VSI-Bench (Yang et al., 2025a) evaluates spatial understanding from indoor-scene videos, requiring models to integrate observations across frames to recover scene layout and object relations. The test split contains 5,130 questions. It covers four numerical tasks—object counting, absolute distance, object size, and room size—and four multiple-choice task groups—relative distance, relative direction, route planning, and appearance order. Following the official protocol, numerical answers use mean relative accuracy averaged over thresholds t from 0.50 to 0.95 in increments of 0.05, counting an answer as correct when its relative error is at most $1 - t$. Multiple-choice answers use exact-match accuracy; relative direction first averages its easy, medium, and hard subsets. The overall score is the unweighted mean over all eight task groups. The numerical and multiple-choice aggregates in Table 3 each average the corresponding four groups. The category means discussed in Sec. 4.2 are computed from unrounded task scores before rounding to one decimal place.

MMSI-Bench. MMSI-Bench (Yang et al., 2026b) evaluates spatial reasoning from multiple images, requiring models to relate objects, scene regions, and camera viewpoints across observations of real-world scenes. We evaluate its 1,000-question test set. The ten fundamental tasks use two images per question, while multi-step reasoning questions use more images. The 11 question types cover appearance and measurable attributes, camera and object motion, positional relations among cameras, objects, and regions, and multi-step reasoning. Each question is multiple-choice. Our evaluator computes exact-match accuracy within each question type and reports the unweighted mean over the 11 types, so each type contributes equally despite differences in sample count.

ViewSpatial. ViewSpatial-Bench (Li et al., 2026a), abbreviated as ViewSpatial in our tables, evaluates spatial understanding under camera and person perspectives, testing whether a model can interpret relations in the specified reference frame. Constructed from ScanNet and MS-COCO, its 5,712-question test set contains five task categories: relative direction and object-view orientation from the camera perspective, the same two tasks from a person's perspective, and relative direction under person-perspective scene simulation. The questions use multiple-choice answers. Our reported score is the unweighted mean of exact-match accuracy over the five categories.

CV-Bench. CV-Bench (Tong et al., 2024) tests visual perception and spatial understanding from single images using 2,638 manually inspected multiple-choice questions derived from ADE20K, COCO, and Omni3D. Its 2D tasks cover spatial relationships and object counting, while its 3D tasks cover depth order and relative distance. This benchmark complements the multi-view evaluations by measuring spatial judgments that can be made from an individual image. Our evaluator computes exact-match accuracy separately over the 2D and 3D questions and reports the unweighted mean of these two scores.

MindCube. MindCube (Yin et al., 2025) evaluates spatial mental modeling from limited views through cognitive mapping, perspective-taking, and mental simulation. We use the 1,050-question MindCube-Tiny subset, comprising 600 *Among*, 250 *Around*, and 200 *Rotation* questions. These settings vary camera motion: moving among objects, moving around objects, and rotating in place, respectively. They require models to combine partial observations and reason about positions, orientations, and hypothetical movements. Our MindCube results in Table 2 use this subset. We report sample-averaged exact-match accuracy on its multiple-choice questions.

Cross-benchmark aggregation. Table 2 reports overall accuracy (%) on MMSI-Bench, ViewSpatial, CV-Bench, and MindCube. Its average is the unweighted mean across the four benchmarks and is reported only when all four scores are available. Missing or unreported entries are marked with "–" in the comparison tables.

SPBench. Our complementary analysis uses SPBench's (Li et al., 2026c) single-image (SI) and multi-view (MV) subsets, containing 1,009 and 319 questions, respectively. SI covers absolute distance, relative direction, relative distance, and object size; MV additionally includes object counting. We retain the evaluator's multiple-choice accuracy and numerical mean relative accuracy (MRA). The score for each subset is the unweighted mean of its task-category metrics, and the overall score is the unweighted mean of SI and MV scores. It is therefore distinct from accuracy pooled over all questions, and is not the simple mean of the Accuracy and MRA columns in Table 4. SPBench is analyzed separately from the four-benchmark average in Table 2. The detailed metrics are reported in Appendix D.1.

Baseline reporting. The comparison tables separate proprietary models, open-source generalist models, and models specialized for spatial intelligence, with model sizes shown in the method names. Spatial baselines include SPAR, SpatialLadder, Spatial-MLLM, VG-LLM, CA-MLLM, and GeoThinker. Unless otherwise stated, VSI-Bench baseline scores follow the values reported by the corresponding papers under the VSI-Bench protocol. These reported-score comparisons should be distinguished from the component ablations, which use the Qwen3-VL-4B backbone.

B.2 Training and Inference Settings

Backbones. The backbone configurations are Qwen2.5-VL-3B, and Qwen3-VL-4B (Bai et al., 2025b;a), with VGGT-1B as the geometry encoder (Wang et al., 2025). Both the semantic vision tower and the geometry encoder are frozen. We optimize the language model and geometry integration modules for one epoch on the 234K spatial and 64K general video instruction mixture described in Sec. 4.1. The populated rows on VSI-Bench and the four additional benchmarks use Qwen2.5-VL-3B and Qwen3-VL-4B.

Optimization. Training uses AdamW with bfloat16 precision, learning rate 10^{-5}, weight decay 0.01, a cosine learning-rate schedule, and a 3% warmup. The effective batch size is 64 across eight NVIDIA A800 GPUs, and the maximum sequence length is 12,800 tokens. Training samples contain 4–8 frames. Video evaluation uniformly samples up to 32 frames and uses greedy decoding with batch size one. The wall-clock training time and number of repeated runs remain to be reported with the final experiments.

Geometry and memory defaults. Unless varied in the sensitivity analysis, we use physical-group thresholds $\delta_x = 0.08$ and $\delta_f = 0.6$, selector temperature $\tau = 1.0$, text-rater temperature $\tau_r = 0.1$, and spatial-prefix ratio $\rho = 0.01$. The remaining grouping and selector settings are given in Appendix C.

C Method Details

C.1 Physical Grouping and Memory Settings

Valid observations and physical grouping. We discard VGGT camera and register tokens and mask non-finite points, non-finite confidences, and degenerate rays. The matching thresholds in Sec. 3.2 are $\delta_x = 0.08$ and $\delta_f = 0.6$. Before constrained union–find, candidate edges are sorted by $d/\delta_x + \frac{1}{2}(1 - \mathbf{z}_{vi}^{\top}\mathbf{z}_{uj})$, where d is their distance in normalized 3D space. Confidence-weighted group features and centroids fall back to unweighted averages when all weights vanish.

Question selection and memory settings. During training, an explicit question mask selects embeddings from the latest user turn, excluding system text, previous turns, answer tokens, and visual placeholders. During generation without this mask, retrieval uses valid prompt tokens after the final vision block, including the chat-format suffix; if that span is empty, it uses the non-visual prompt. Both paths retain at most the latest 256 eligible tokens. The selector width is $d_s = 256$, the affinity temperature is $\tau = 1.0$, and the text-rater backward temperature is $\tau_r = 0.1$. We use consensus weight $\beta = 0.05$, patch-confidence weight $\gamma = 0.25$, and spatial-prefix ratio $\rho = 0.01$.

Geometry attention initialization. The learnable gates η in Eq. 4 and g_ℓ in Eq. 5 are initialized to zero, making the corresponding attention residuals zero at the start of fine-tuning.

C.2 Spatial Prefix Group Encoder

The pointwise group encoder conditions each canonical feature on

$$\mathbf{m}_r = [\overline{\mathbf{x}}_r,\, C_r,\, n_r/V,\, \log(1+|r|)/\log(1+V),\, D_r], \tag{6}$$

where $|r|$ is the number of observations and D_r is their mean cosine dispersion around the canonical feature. A LayerScale residual MLP produces

$$\widehat{\mathbf{g}}_r = \overline{\mathbf{g}}_r + \boldsymbol{\lambda} \odot \mathbf{W}_o\, \mathrm{GELU}(\mathbf{W}_f\, \mathrm{LN}(\overline{\mathbf{g}}_r) + \mathbf{W}_m \mathbf{m}_r)\,. \tag{7}$$

These refined features are pooled into the spatial anchors described in Sec. 3.3.

D Supplementary Experiments

This section provides detailed SPBench metrics, the component-ablation reporting protocol, an analysis of token selection and fusion, and analyses of parameter sensitivity and inference cost.

D.1 Detailed Results on SPBench

Table 4 separates multiple-choice accuracy from numerical mean relative accuracy (MRA) for single-image (SI) and spatial-scene (SS) evaluation. Following the terminology in Sec. 4.3, SS denotes SPBench's multi-view subset, defined in Appendix B.1. This breakdown complements Fig. 3 by showing whether the observed gains extend across both answer formats. The aggregate subset scores average task-category metrics and are therefore not the simple mean of the accuracy and MRA columns.

Table 4: **SPBench results by input setting and answer format.** SI uses single images; SS uses multiple views of a shared scene. Accuracy evaluates multiple-choice answers, and MRA evaluates numerical answers, both on a 0–100 scale. Qwen2.5-3B and Qwen3-4B denote the Qwen2.5-VL-3B and Qwen3-VL-4B backbones, respectively.

	SI		SS	
Model	Accuracy	MRA	Accuracy	MRA
Qwen2.5-VL-3B	41.81	30.16	42.62	32.02
CA-MLLM	52.39	28.38	49.18	60.70
VG-LLM	50.13	49.56	47.54	65.50
GeoThinker$_{\text{Qwen2.5-3B}}$	52.14	48.20	50.82	61.98
GS-MLLM$_{\text{Qwen2.5-3B}}$	51.39	48.71	52.46	66.94
GeoThinker$_{\text{Qwen3-4B}}$	58.44	51.32	57.38	64.22
GS-MLLM$_{\text{Qwen3-4B}}$	57.43	51.44	62.30	67.13

Spatial-scene gains span both answer formats. Compared with GeoThinker using the same backbone, GS-MLLM improves SS accuracy and MRA by **1.64** and **4.96 points**, respectively, with Qwen2.5-VL-3B. With Qwen3-VL-4B, the corresponding gains are **4.92** and **2.91 points**, reaching 62.30 accuracy and 67.13 MRA, the highest values in both SS columns. The SS advantage therefore spans both categorical judgments and numerical estimation across the two backbones.

Single-image results remain mixed. On SI, GS-MLLM has lower accuracy than the corresponding GeoThinker variant with both backbones: 51.39 versus 52.14 with Qwen2.5-VL-3B, and 57.43 versus 58.44 with Qwen3-VL-4B. MRA increases only slightly, from 48.20 to 48.71 and from 51.32 to 51.44, respectively. This contrast locates the consistent improvement in the spatial-scene setting, where multiple observations describe a shared environment. Because SI and SS contain different questions and task compositions, their score differences do not isolate the effect of adding views to the same reasoning problem.

D.2 Component Ablation Reporting

Section 4.4 explains the role of each variant in Table 3. All variants use Qwen3-VL-4B. The full-model reference is taken from Table 1. The overall score averages all eight VSI-Bench task groups; the numerical and multiple-choice aggregates each average four groups as specified in Appendix B.1. Reporting both aggregates prevents gains in categorical reasoning from obscuring changes in metric estimation. Scores are reported on a 0–100 scale and rounded to two decimal places after aggregation.

D.3 Token Selection and Fusion Strategies

We conduct two supplementary comparisons: token selection under a shared fusion interface on MindCube-Tiny, and fusion strategies on VSI-Bench. The supplementary interface appends Spatial and Semantic tokens after the QA tokens in the language-model sequence. In GS-MLLM, retrieved geometry enters Pixel-token states through coordinate-grounded, view-aligned cross-attention (Sec. 3.4), while a global spatial prefix precedes the first vision-start token (Sec. 3.3).

Token selection on MindCube-Tiny. Table 5 compares three ways to supply VGGT features under post-QA concatenation: (i) dense features, unpruned at training and inference; (ii) object-mask selection, using Qwen3-VL-8B to identify image subjects and SAM masks to retain their VGGT features; and (iii) question-conditioned selection with 197 or 301 tokens per view. Evaluation uses the 1,050-question MindCube-Tiny subset in Appendix B.1: 250 Around, 200 Rotation, and 600 Among questions. Overall accuracy averages all questions, so Among contributes most heavily; no translation questions are evaluated.

Table 5: **Token selection under concatenation-based fusion on MindCube-Tiny.** Spatial and Semantic tokens are appended after the QA tokens in every row. Scores are exact-match accuracy (%); the best score in each column is bold. The budget counts selected geometry tokens per view; "–" denotes an unreported count.

Selection	Budget/view	Around	Rotation	Among	Overall
Dense VGGT	All	66.40	47.50	60.33	59.33
Object masks	–	67.20	**51.50**	63.50	62.10
Question-conditioned	197	74.00	45.50	**69.33**	**65.90**
Question-conditioned	301	**75.20**	46.50	67.83	65.52

Object-mask selection improves all three settings over dense features, raising overall accuracy from 59.33 to 62.10. Question-conditioned selection with 197 tokens achieves the highest overall accuracy, 65.90, with 69 more correct answers than dense features and 40 more than object masks. It improves Around (74.00 vs. 67.20) and Among (69.33 vs. 63.50) over object masks, but scores lower on Rotation (45.50 vs. 51.50), where object masks perform best.

Increasing the budget from 197 to 301 tokens improves Around by 1.20 points and Rotation by 1.00 point, but lowers Among by 1.50 points. Overall accuracy decreases from 65.90 to 65.52, corresponding to four fewer correct answers (688 vs. 692). The 197-token configuration uses 34.6% fewer selected tokens and achieves a slightly higher aggregate score. This comparison supports the smaller budget among the tested settings, without establishing statistical superiority or a runtime benefit.

Fusion strategies on VSI-Bench. We next compare the 197-token post-QA concatenation configuration with GS-MLLM on VSI-Bench. Both use Qwen2.5-VL-3B and the training recipe in Appendix B.2. Table 6 reports the eight task-group scores and their numerical, multiple-choice, and overall means under the standard VSI-Bench protocol.

Table 6: **Fusion configurations on VSI-Bench with Qwen2.5-VL-3B and the same training recipe.** Post-QA concatenation uses question-conditioned selection with 197 geometry tokens per view. GS-MLLM results follow Table 1, with means computed from unrounded scores. Concatenation task-family means use the reported task scores. All scores use a 0–100 scale, with the higher score in each row in bold.

Metric	Post-QA concatenation	GS-MLLM
Object counting	30.7	**68.1**
Absolute distance	31.3	**36.2**
Object size	40.1	**57.9**
Room size	28.6	**60.9**
Relative distance	38.5	**44.9**
Relative direction	42.7	**47.0**
Route planning	33.3	**35.1**
Appearance order	26.7	**43.2**
Numerical mean	32.7	**55.8**
Multiple-choice mean	35.3	**42.6**
Overall	34.0	**49.2**

GS-MLLM scores 49.2 overall, compared with 34.0 for post-QA concatenation, a 15.2-point gain. The advantage is larger on numerical tasks (23.1 points) than on multiple-choice tasks (7.3 points). GS-MLLM improves all eight task groups, with the largest gains in object counting (37.4 points) and room size (32.3 points). These results favor the main integration design across both answer formats. The comparison evaluates complete interfaces and does not isolate token placement from memory construction or geometry attention.

D.4 Parameter Sensitivity and Inference Cost

We examine the geometric distance threshold δ_x, feature-similarity threshold δ_f, and spatial-prefix ratio ρ with Qwen3-VL-4B on VSI-Bench. Table 7 reports performance and memory sizes for settings around the defaults $(\delta_x, \delta_f, \rho) = (0.08, 0.6, 0.01)$.

Table 7: **Sensitivity of geometry construction and memory size.** Qwen3-VL-4B on VSI-Bench; default settings are underlined.

Parameter	Setting	VSI Avg.	Mean groups	Mean prefix tokens
δ_x	0.04	54.23	20,726.15	207.72
	<u>0.08</u>	54.65	20,492.00	205.44
	0.12	54.66	20,458.05	205.09
δ_f	0.4	54.43	20,078.93	201.35
	<u>0.6</u>	54.65	20,492.00	205.44
	0.8	54.33	22,903.12	229.56
ρ	0.005	54.01	20,492.00	102.95
	<u>0.01</u>	54.65	20,492.00	205.44
	0.03	54.60	20,492.00	615.30

VSI Avg. averages the eight task-group scores over 5,130 questions. The default-model score is from Table 1. Memory counts average 5,136 forwards, including distributed padding. For $G \geq 1$ valid groups, the prefix contains $\lceil \rho G \rceil$ tokens; the mean is computed after rounding each forward's token count up.

Grouping thresholds. The distance and feature thresholds control which cross-view observations can form a physical group. At $\delta_x = 0.04$ and 0.12, the model scores 54.23 and 54.66, with mean group counts of 20,726.15 and 20,458.05, respectively. Increasing δ_f from 0.4 to 0.8 imposes a stricter feature-matching criterion and increases the mean group count from 20,078.93 to 22,903.12, while the corresponding scores are 54.43 and 54.33. These results show limited score variation over the tested thresholds, even as the feature criterion changes the size of the grouped representation.

Spatial-prefix budget. The ratio ρ controls how many prefix tokens summarize the physical groups. Increasing ρ from 0.005 to 0.03 leaves grouping unchanged and expands the mean prefix length from 102.95 to 615.30 tokens; the corresponding scores are 54.01 and 54.60. The broader budget sweep in Fig. 4 (left) reaches its highest score, 54.65, at $\rho = 0.01$, using approximately 205 tokens. Larger budgets increase evaluation time without improving this score, supporting the 1% prefix as a compact operating point among the tested settings.

Evaluation-time measurement. The prefix sweep uses eight GPUs with batch size one and greedy decoding. Scores average the eight VSI-Bench task groups over 5,130 unique questions. Evaluation time per case is the total wall time divided by 642 cases per GPU, including distributed padding to 5,136 forwards. This measure includes model loading, preprocessing, generation, scoring, and synchronization. Each budget has one timing observation, and the fitted line summarizes the observed trend. For the model comparison in Fig. 4 (right), reported times are normalized by Qwen2.5-VL-3B's 2.1325 s per case.

E Qualitative Case Details

Generation-time records. The two cases in Figure 5 use generation-time traces from the Qwen3-VL-4B-based GS-MLLM checkpoint `HardQ-PatchKV-Prefix1pct_v1`, with greedy decoding, $\rho = 0.01$, 32 uniformly sampled frames. Answers, selection masks, and pooling assignments come from the same run for each case. The direction example is VSI-Bench case 1662 from ScanNet scene `scene0629_01`; the size example is case 898 from ScanNet++ scene `acd95847c5`. Both predictions match the annotations. The direction answer is recorded as option B and displayed as its label, left; the size answer is recorded as 200, with the question's unit (cm) added for readability. The parenthetical list of dimensions is omitted from the displayed size question for readability.

Spatial regions. Each color identifies one spatial-prefix anchor and is fixed across frames within a case. All valid patches in each displayed frame are colored using their recorded physical-group and prefix-anchor assignments, without re-clustering or manual region filtering. The colors indicate pooling membership, not object categories or attention weights.

Semantic-guided selection. The red image patches show the recorded question-conditioned hard retrieval. All selected patches in each displayed frame are shown. This image overlay visualizes the result of Semantic-token guidance; the selected text raters themselves are given below.

Semantic selection. The highlights are read from the recorded binary text-rater mask m_ℓ in Eq. 2. Token boundaries, ordering, and repeated occurrences are preserved. The complete captured conditioning streams are shown below, including the video preamble, answer instructions, and chat-format suffix. Yellow marks selected raters; gray marks unselected tokens; ↩ denotes a recorded newline. This mask controls geometry scoring; it does not remove tokens from the language-model input.

Case 1662: complete Semantic selection (29/59)

These are frames of a video.↩ If I am standing by the bed and facing the chair, is the mirror to the left or the right of the chair?↩ Options:↩ A. right↩ B. left↩ Answer with the option's letter from the given choices directly.

Case 898: complete Semantic selection (26/49)

These are frames of a video.↩ What is the length of the longest dimension (length, width, or height) of the white board, measured in centimeters?↩ Please answer the question using a single word or phrase.

E.1 Additional Visual Case Studies

The following eight examples include six correct answers across six task types and two failures, collected under the same generation settings described above. Each block contains eight chronologically ordered input images per row, sampled uniformly from the 32 recorded inputs. The upper row colors all spatial-prefix areas; the lower row marks Semantic-guided retrieved patches in red. The two rows use identical frames, and a spatial area retains its color within each case. The Answer field gives the ground truth; multiple-choice answers and predictions are displayed as option text. These examples illustrate model behavior and do not estimate the frequency of success or failure.

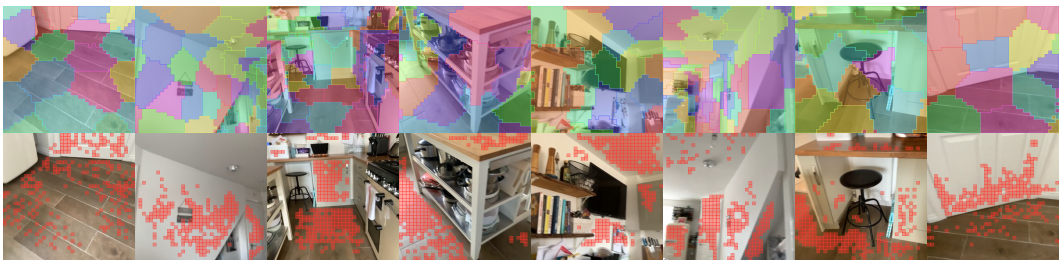
Case 1: Object counting **Correct**



Question: How many bookshelf(s) are in this room?

Answer: 2
Prediction: 2

Case 2: Object size **Correct**



Question: What is the length of the longest dimension (length, width, or height) of the stove, measured in centimeters?

Answer: 100 cm
Prediction: 100 cm

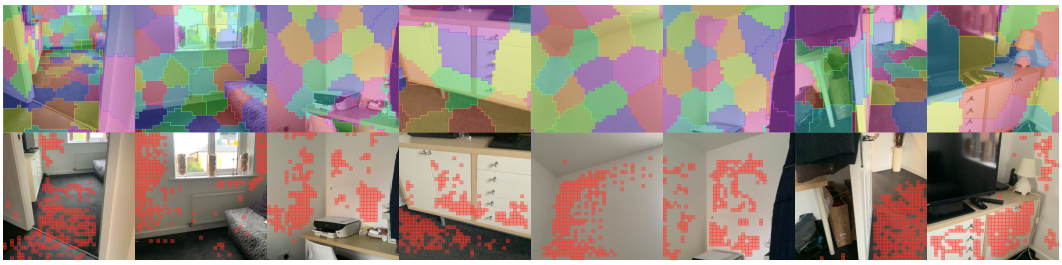
Case 3: Relative distance **Correct**



Question: Measuring from the closest point of each object, which of these objects (chair, stool, stove, sofa) is the closest to the tv?

Answer: chair
Prediction: chair

Case 4: Relative direction **Correct**



Question: If I am standing by the tv and facing the sofa, is the table to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

Answer: front-right
Prediction: front-right

Case 5: Appearance order **Correct**



Question: What will be the first-time appearance order of the following categories in the video: toilet, bed, basket, table?

Answer: toilet, basket, bed, table
Prediction: toilet, basket, bed, table

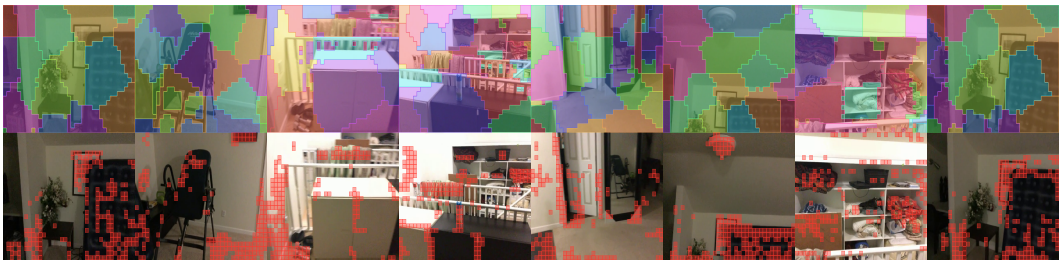
Case 6: Route planning **Correct**



Question: You are a robot beginning at the refrigerator and facing the refrigerator. You want to navigate to the Christmas tree. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the stove. 3. [please fill in] 4. Go forward until the Christmas tree. You have reached the final destination.

Answer: Turn Left, Turn Left
Prediction: Turn Left, Turn Left

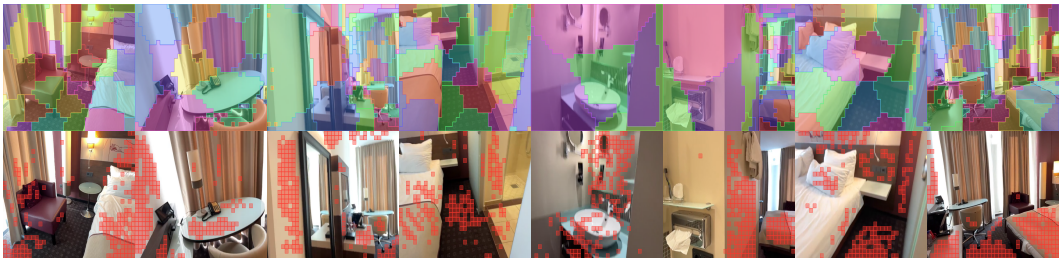
Case 7: Relative direction **Failure**



Question: If I am standing by the bed and facing the mirror, is the chair to the left or the right of the mirror?

Answer: right
Prediction: left

Case 8: Route planning **Failure**



Question: You are a robot beginning at the chair and facing the table. You want to navigate to the apartment door. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the door. You have reached the final destination.

Answer: Turn Back
Prediction: Turn Right