TOKEN-ADAPTIVE ROUTING OF MULTI-LEVEL GEOMETRY FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

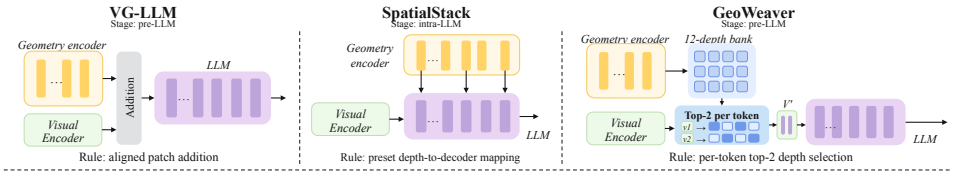
Spatio-temporal reasoning in vision-language models requires visual representations that preserve physical geometry rather than merely semantic appearance. Recent multimodal models incorporate geometric information through structural branches, 3D-aware supervision, reasoning-stage fusion, or long-horizon memory. While these approaches demonstrate the importance of geometry for spatial intelligence, many combine geometric features with visual representations through shared or stage-wise rules. This leaves a finer-grained question: which levels of geometric evidence should each visual token use for its spatial context? To address this question, we introduce **GeoWeaver**, a token-adaptive geometric grounding framework that selects multi-level evidence before language decoding. GeoWeaver constructs a multi-level geometry bank from a frozen geometry encoder, and each visual token sparsely selects evidence from this bank. The selected evidence is incorporated into visual tokens via a residual grounding operation prior to language modeling, yielding geometry-grounded representations for downstream reasoning. Across six reported spatial reasoning benchmarks, GeoWeaver attains a 64.8 mean. In a matched 32-frame ReVSI comparison, it exceeds SpatialStack by 2.9 points, while its four-task general multimodal mean improves on the base model (68.98 versus 68.41). On VSI-Bench, token-adaptive allocation reaches 72.7, compared with 71.30 for global layer weighting and 70.44 for uniform averaging, supporting the value of selecting geometric evidence at token granularity. Code and models will be released.

1 INTRODUCTION

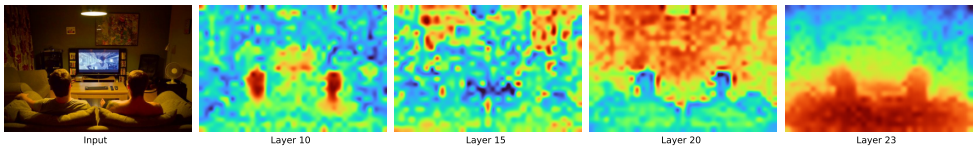
Understanding and reasoning about physical space is a fundamental capability for intelligent systems that must perceive, communicate, and act in the real world. Recent advances in large vision-language models (VLMs) significantly improve general visual understanding, enabling strong performance in image and video question answering, instruction follow-up, and multimodal reasoning. However, reliable spatio-temporal reasoning remains a major challenge. Tasks involving relative distance estimation, object layout understanding, temporal tracking of spatial relations, or grounding motion and direction in dynamic scenes continue to expose systematic failures that cannot be explained by language reasoning alone.

To address this limitation, recent studies increasingly incorporate spatial information into multimodal models, progressing from perception-side enhancement toward reasoning-oriented spatial modeling. As shown in Figure 1, early efforts strengthen visual representation by introducing explicit spatial or structural branches, allowing geometry-aware cues to complement semantics-dominated visual features (Wu et al., 2025a; Zheng et al., 2025a). Building on this direction, representation-level approaches further adopt 3D-aware supervision or geometry-oriented objectives to improve the structural sensitivity of visual tokens, suggesting that spatial reasoning depends not only on input design but also on the geometry encoded in intermediate representations. More recent geometry-enhanced MLLMs investigate how structural cues interact with language reasoning, either by injecting geometric features into the multimodal reasoning stack or by aligning spatial signals with hierarchical reasoning stages (Fan et al., 2025; Li et al., 2025d). Beyond single-instance reasoning, long-horizon spatial cognition extends the problem to continuous video streams, where predictive sensing, memory organization, and event-centric modeling support temporally persistent spatial understanding (Yang

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.



(a) Representative integration stages and allocation rules.



(b) Depth-dependent VGGT feature responses.

Figure 1: **Geometry integration stage and evidence allocation.** (a) Representative designs, from left to right: VG-LLM (Zheng et al., 2025a) uses aligned patch addition before the LLM; SpatialStack (Zhang et al., 2026a) maps geometry depths to decoder layers; GeoWeaver selects encoder depths per visual token before the LLM. VG-LLM and GeoWeaver are both pre-LLM; their allocation rules differ. (b) Selected VGGT layers show different spatial responses, motivating a multi-level evidence bank from which visual tokens can select evidence.

et al., 2025d). These approaches establish the value of geometry at several integration stages; they also raise a more specific question of how each visual token should select evidence across geometry-encoder depths.

Figure 1b shows that geometry-encoder depths respond differently to the same scene. Visual tokens likewise represent distinct image regions, so a single depth mixture shared across tokens need not provide the most useful evidence at every location. We therefore study whether selecting geometry depths separately for each visual token can produce stronger *geometry-grounded* representations for spatio-temporal reasoning.

Motivated by this observation, we formulate *token-adaptive geometric grounding*: Each visual token predicts routing weights over geometry-encoder depths and aggregates a compact subset of spatially aligned features. Specifically, a frozen geometry encoder constructs the bank from the same input frames. Each visual token queries its aligned bank entries and retrieves a compact mixture of geometric abstractions suited to its location. Residual grounding incorporates that mixture into the visual representation, giving the language model tokens that carry both semantic and selected geometric evidence. The pre-decoder interface is where we apply this allocation; the distinguishing operation is the token-specific choice across geometry depths.

We instantiate this formulation as **GeoWeaver** and evaluate its effectiveness across spatial reasoning benchmarks. Allocation ablations examine the contribution of token-specific depth selection, while a cross-view correspondence probe assesses the geometric information carried by the grounded visual tokens. This lets different regions draw on different geometric evidence instead of receiving one shared depth mixture. GeoWeaver performs strongly across spatial reasoning benchmarks, and an allocation ablation directly compares uniform, global, and token-adaptive depth selection. A cross-view correspondence probe further examines the resulting visual representations. The main contributions of this work are: The main contributions of this work are:

- We propose **GeoWeaver**, a token-adaptive geometric grounding framework for spatio-temporal reasoning. GeoWeaver replaces a shared geometry-depth mixture with token-specific evidence allocation, allowing different image regions to draw on different geometric abstractions.
- We develop a sparse cross-depth routing mechanism over a multi-level geometry bank. Each visual token selects and combines spatially aligned features from a compact subset of geometry-encoder layers. Residual grounding incorporates the selected evidence into native visual tokens before language decoding, without extending the visual token sequence.
- Extensive experiments across spatial reasoning benchmarks and two model scales demonstrate the strong performance of GeoWeaver. Allocation ablations support the effectiveness

of token-specific depth selection, while cross-view correspondence analysis reveals improved geometric correspondence in the grounded visual representations.

2 Related Work

Spatial Intelligence in MLLMs. Recent large multimodal models (MLLMs) (Liu et al., 2023; Bai et al., 2025a; Wang et al., 2025e; Liu et al., 2026) show strong general-purpose visual understanding, but spatial reasoning in physical environments remains challenging. Recent benchmarks, including VSI-Bench (Yang et al., 2025a), SPAR (Zhang et al., 2025), BLINK (Fu et al., 2024), and Cambrian-S (Yang et al., 2025d), evaluate this limitation through tasks involving depth, viewpoint, topology, and spatiotemporal relations (OpenAI, 2025b; Gemini, 2025; Zhou et al., 2025). These studies indicate that spatial intelligence is not merely a difficult subset of generic multimodal reasoning. Instead, it requires visual representations that preserve geometry-sensitive evidence beyond semantics-dominated appearance cues. Our work follows this diagnosis, but focuses on the representation before reasoning: whether visual tokens are geometrically grounded when they enter the language model.

Data Scaling and Spatial Reasoning Supervision. Much effort improves spatial reasoning through data construction and specialized optimization. SpatialVLM (Chen et al., 2024), SPAR (Zhang et al., 2025), and Cambrian-S (Yang et al., 2025d) enlarge the supervision space with synthetic 3D scenes, physically grounded tasks, and long-horizon spatial cognition settings. Furthermore, VST (Yang et al., 2025c), SpaceR (Ouyang et al., 2025a), MindCube (Yin et al., 2025), SpatialLadder (Li et al., 2025d), and Spatial-SSRL (Liu et al., 2025c) exploit structured tuning, reinforcement learning, and verifiable rewards to strengthen spatial reasoning behaviors. SenseNova-SI (Cai et al., 2025a) further suggests that scaling text-based chain-of-thought alone brings limited gains, highlighting that spatial reasoning depends on more than language-side deliberation. These methods provide stronger supervision and optimization signals, while our work addresses a complementary question: how the visual representation should be organized before such reasoning takes place.

Geometry-Augmented Multimodal Modeling. A closely related direction incorporates explicit geometric priors into MLLMs. Some methods introduce external 3D inputs, such as point clouds or RGB-D signals, to connect language models with spatial structure (Du et al., 2024; Zhu et al., 2025a; Zheng et al., 2025b). More recently, feed-forward geometry encoders, including DUST3R (Wang et al., 2024; Leroy et al., 2024), CUT3R (Wang et al., 2025b), and VGGT (Wang et al., 2025a), provide dense geometric estimates from multi-view images or videos. Building on these encoders, Spatial-MLLM (Wu et al., 2025a), VG-LLM (Zheng et al., 2025a), VLM-3R (Fan et al., 2025; 2026), SSR (Liu et al., 2025a), and related models (Li et al., 2026) introduce geometry through structural branches, patch-level fusion, cross-attention, or rationale-guided integration. Some, including VG-LLM, already fuse geometric information before the LLM; VLM-3R also uses visual-token-conditioned interaction. GeoWeaver instead focuses on a distinct allocation rule: each visual token sparsely selects from aligned geometry-encoder depths before language decoding.

Multimodal Chain-of-Thought and Visual Intermediate Reasoning. Recent unified multimodal architectures (Team, 2024; Deng et al., 2025; Jiao et al., 2025) also motivate a line of work that externalizes intermediate reasoning through visual generation or transformation. For example, some methods render maze trajectories (Wu et al., 2024), imagine temporal transitions (Li et al., 2025b), or draw intermediate 2D geometric cues (Li et al., 2025a; Gu et al., 2025). Such visual intermediates help when the required transformation is difficult to express purely in text. However, spatio-temporal reasoning in real 3D scenes also requires dense geometric structure, viewpoint consistency, and continuous physical relations to remain available inside the representation. GeoWeaver does not externalize reasoning into generated visual artifacts; instead, it improves the visual tokens consumed by the language model through pre-reasoning geometric grounding.

3 Method

Motivated by the different responses of geometry-encoder depths, GeoWeaver uses token-adaptive selection from a multi-level geometry bank to prepare visual representations for spatio-temporal reasoning. Grounding takes place before language decoding so that the selected evidence is incorporated into the tokens consumed by the language model. We first introduce the semantic and geometric

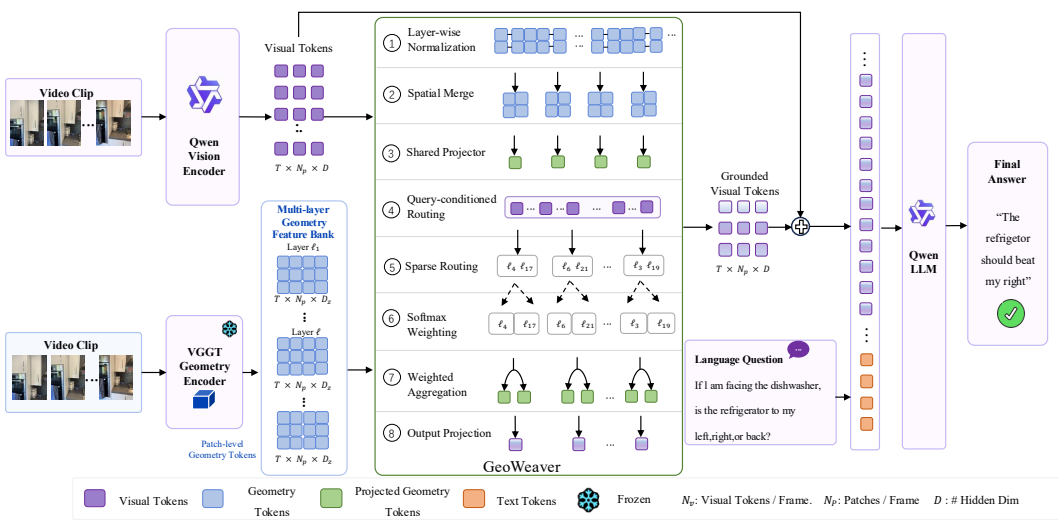


Figure 2: **Overview of GeoWeaver.** GeoWeaver builds a multi-layer geometry bank with a frozen VGGT encoder. Each visual token queries its aligned bank entries, selects two geometry depths by default, and receives their evidence through a residual update before entering the Qwen LLM. The resulting tokens combine semantic visual content with token-selected geometry for spatio-temporal reasoning.

feature extraction branches, then describe token-adaptive geometric grounding, and finally present language reasoning over grounded visual tokens and the training objective.

3.1 Overview

Given a sequence of input frames $\{\mathbf{I}_k \in \mathbb{R}^{H\times W\times 3}\}_{k=1}^{K}$ and a language instruction q, our framework consists of three stages: (1) a semantic visual branch extracts visual tokens from the base VLM, (2) a frozen geometry branch constructs a multi-level geometry bank from the same frames, and (3) a token-adaptive grounding stage allocates geometric evidence to visual tokens before language decoding.

Visual branch. We build our method on top of a pretrained multimodal LLM. In our implementation, we use a Qwen3.5 VLM backbone, although the framework is model-agnostic. Given K input frames, the native vision encoder produces patch tokens, which are further processed by the model's built-in token merger. Let

$$\mathbf{V} \in \mathbb{R}^{(KN_p)\times D} \tag{1}$$

denote the resulting visual tokens, where N_p is the number of merged spatial tokens per frame and D is the hidden dimension expected by the language model.

Geometry branch. In parallel, we use a frozen geometry encoder to extract explicit geometric priors from the same input frames. We adopt VGGT (Wang et al., 2025a) as the geometry encoder. Rather than using only its final representation, we collect a contiguous set of deeper hidden states to form a *multi-level geometry bank*. Specifically, for each selected layer $l \in \mathcal{S}$, we take the corresponding patch-token hidden state, discard non-patch tokens such as camera or register tokens, and obtain

$$\mathbf{Z}^{(l)} \in \mathbb{R}^{(KN_v)\times D_{\text{geo}}}, \qquad l \in \mathcal{S}, \tag{2}$$

where N_v is the number of raw geometry tokens per frame and D_{geo} is the geometry feature dimension. In practice, we use the latter half of the geometry encoder layers (layers 12–23 in a 24-layer VGGT), which provide rich structural abstractions while preserving spatial detail.

3.2 Token-Adaptive Geometric Grounding

The core of our framework is a token-adaptive grounding process that injects geometric evidence into visual tokens before language decoding. The process consists of four steps: per-layer normalization, spatial alignment, shared projection, and query-conditioned sparse routing.

Per-layer normalization and spatial alignment. As shown in Fig. 1, geometry features from different layers exhibit distinct response patterns and feature distributions. We therefore normalize each selected geometry layer independently:

$$\hat{\mathbf{Z}}^{(l)} = \mathrm{LN}\Big(\mathbf{Z}^{(l)}\Big), \qquad l \in \mathcal{S}. \tag{3}$$

Because the geometry branch and base VLM operate at different spatial scales, direct token-wise grounding is not feasible without alignment. To match the visual token resolution, we apply a 2×2 spatial merge to the normalized geometry tokens, followed by a shared projector:

$$\mathbf{G}^{(l)} = \phi\Big(\mathrm{Merge}_{2\times 2}\Big(\hat{\mathbf{Z}}^{(l)}\Big)\Big), \qquad \mathbf{G}^{(l)} \in \mathbb{R}^{(KN_p)\times D}, \tag{4}$$

where $\phi(\cdot)$ is a lightweight MLP shared across all geometry layers. After this step, all geometry layers are aligned to the same token resolution and hidden dimension as the visual tokens in Eq. 1.

Query-conditioned geometric evidence allocation. A key question is which geometric abstraction should be assigned to each visual token. Instead of exposing all tokens uniformly to every geometry layer, we let each visual token query the multi-level geometry bank. For the i-th visual token $\mathbf{v}_i \in \mathbb{R}^D$, a lightweight router predicts its preference over the selected geometry layers:

$$\mathbf{r}_i = f_{\text{router}}(\mathbf{v}_i), \qquad \mathbf{r}_i \in \mathbb{R}^{|\mathcal{S}|}. \tag{5}$$

This design reflects our hypothesis that different visual regions require different geometric abstractions: local boundaries, depth discontinuities, and occlusion patterns need not depend on the same evidence as global layout or relational structure.

Compact geometric grounding. Not all geometry layers provide equally useful evidence for a given visual token. Aggregating the entire geometry bank introduces both relevant and weakly related structural cues, diluting token-specific grounding. We therefore adopt a compact evidence selection strategy that retains only the geometry layers with the highest token-conditioned evidence scores. Given routing logits $\mathbf{r}_i$, we compute layer weights as

$$\boldsymbol{\alpha}_i = \mathrm{Softmax}(\mathcal{S}_{TopK}(\mathbf{r}_i, K)), \tag{6}$$

where $\mathcal{S}_{TopK}(\cdot, K)$ keeps the K most relevant geometry layers and masks the remaining entries before normalization.

Given the aligned geometry bank $\{\mathbf{G}^{(l)}\}_{l\in\mathcal{S}}$, the geometric evidence assigned to token i is

$$\mathbf{g}_i = \sum_{l\in\mathcal{S}} \alpha_i^{(l)} \mathbf{G}_i^{(l)}, \tag{7}$$

where $\mathbf{G}_i^{(l)}$ denotes the geometry token aligned with $\mathbf{v}_i$ in layer l. This operation aggregates multi-layer geometry into compact token-specific evidence for residual grounding.

Residual grounding before reasoning. Finally, we inject the allocated geometric evidence into each visual token through a residual grounding operation:

$$\mathbf{v}_i' = \mathbf{v}_i + \mathbf{W}_o \mathbf{g}_i, \tag{8}$$

where $\mathbf{W}_o \in \mathbb{R}^{D\times D}$ is a projection matrix. We initialize $\mathbf{W}_o$ to zero so that the grounding branch begins as an identity mapping and gradually learns to reshape the visual representation during finetuning. Collecting all grounded visual tokens yields

$$\mathbf{V}' = [\mathbf{v}_1'; \ldots; \mathbf{v}_{KN_p}'] \in \mathbb{R}^{(KN_p)\times D}. \tag{9}$$

Together with token-conditioned routing over the multi-level bank, this residual update forms our grounding mechanism. The output is a set of *geometry-grounded visual tokens* in which each location carries its selected geometric evidence into language decoding.

3.3 Language Reasoning over Geometry-Grounded Tokens

After grounding, the visual tokens are concatenated with the text tokens derived from the instruction:

$$\mathbf{H}_0 = [\mathbf{V}'; \mathbf{T}], \tag{10}$$

Table 1: **Comparison on VSI-Bench (Yang et al., 2025a).** Under EASI (Cai et al., 2025b), numerical answers use MRA, multiple-choice answers use accuracy, and Avg. is the unweighted mean of eight category scores before rounding. Best and second-best open-source results are **bolded** 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
Human	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
Random Choice(Frequency)	35.5	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Proprietary Models									
Seed-1.6 (Team, 2025)	49.9	43.5	34.4	66.1	52.8	55.1	35.7	44.3	68.0
Gemini-3-Pro (Gemini, 2025)	52.5	38.0	37.8	72.7	44.1	59.9	55.7	45.9	66.0
Grok-4 (xAI, 2025)	47.9	37.2	33.0	60.8	45.4	53.1	39.7	47.4	66.8
GPT-5 (OpenAI, 2025b)	55.0	53.3	34.5	73.3	47.5	63.7	48.7	50.3	68.9
Open-source General Models									
Bagel-7B-MoT (Deng et al., 2025)	31.4	30.1	29.2	35.5	25.8	34.9	41.4	30.4	24.1
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	32.3	32.9	18.2	43.9	31.7	38.0	37.4	28.4	28.0
Qwen3-VL-8B-Instruct (Cloud), 2025)	57.9	67.6	47.0	76.3	61.9	58.0	51.0	35.0	66.3
InternVL3-8B (Zhu et al., 2025b)	42.1	66.0	34.9	43.6	47.5	48.0	39.3	26.3	31.4
Open-source Spatial Intelligence Models									
MindCube-3B (Yin et al., 2025)	17.2	12.9	22.8	4.3	23.5	20.3	15.7	16.0	22.5
SpatialLadder-3B (Li et al., 2025d)	44.9	62.2	35.4	62.0	41.4	45.6	46.5	27.3	38.5
Spatial-MLLM-4B (Wu et al., 2025a)	46.3	66.7	38.1	63.6	35.5	40.4	48.2	33.0	44.3
SpaceR-7B (Ouyang et al., 2025a)	41.5	44.5	24.7	53.5	37.3	41.9	46.1	29.3	54.8
VLM-3R-7B (Fan et al., 2025)	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
ViLaSR-7B (Wu et al., 2025b)	44.6	58.1	33.9	61.4	28.9	45.1	46.5	29.9	53.2
VST-7B-SFT (Yang et al., 2025c)	55.5	68.8	37.3	74.5	62.2	55.2	48.7	41.8	55.5
Cambrian-S-7B (Yang et al., 2025d)	62.9	68.2	45.8	72.5	67.6	66.8	69.6	39.2	73.8
SpatialStack-5B (Zhang et al., 2026a)	67.5	71.0	55.6	69.1	68.2	67.3	84.1	41.2	83.5
Ours									
GeoWeaver-5B	72.7 (+5.2)	83.2 (+12.2)	**61.8** (+6.2)	76.8 (+7.7)	62.4	72.8 (+5.5)	83.5	**56.3** (+15.1)	**84.7** (+1.2)
GeoWeaver-10B	**75.05** (+7.6)	**84.3** (+13.3)	58.8 (+3.2)	**83.8** (+14.7)	**68.8** (+0.6)	**73.5** (+6.2)	**92.0** (+7.9)	55.0 (+13.8)	84.3 (+0.8)

where $\mathbf{T} \in \mathbb{R}^{M \times D}$ denotes the text-token embeddings. The resulting multimodal sequence is processed by the LLM decoder:

$$\mathbf{H}_L^{\text{llm}} = f_L^{\text{llm}}\Big(f_{L-1}^{\text{llm}}\big(\cdots f_1^{\text{llm}}(\mathbf{H}_0)\big)\Big), \tag{11}$$

where $f_j^{\text{llm}}(\cdot)$ denotes the j-th decoder block and $\mathbf{H}_L^{\text{llm}}$ the final hidden states. Grounding at this interface presents the decoder with visual tokens already enriched by token-selected geometric evidence for relational and temporal reasoning.

3.4 Optimization

We train the model using the standard next-token prediction objective (Cloud), 2025):

$$\mathcal{L}_{\text{ce}}(\theta) = -\sum_{t=1}^{|o|} \log P_\theta\big(o_t \mid o_{<t}, q, \{\mathbf{I}_k\}_{k=1}^{K}\big), \tag{12}$$

where o_t denotes the t-th target answer token. During instruction tuning, we keep the geometry encoder and native vision encoder frozen and fully optimize the grounding module. For the language model, the 5B configuration uses full fine-tuning and the 10B configuration uses LoRA. No auxiliary geometric reconstruction or contrastive objectives are introduced; instead, the model learns to allocate and exploit geometric evidence through end-to-end supervision on downstream spatio-temporal reasoning tasks.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate GeoWeaver on both specialized spatial reasoning benchmarks and general multimodal evaluation sets. For spatial reasoning, we report results on **VSI-Bench** (Yang et al., 2025a), **ReVSI** (Zhang et al., 2026b), **SPAR-Bench** (Zhang et al., 2025), **ViewSpatial** (Li et al., 2025c), **BLINK** (Fu et al., 2024), **3DSRBench** (Ma et al., 2025), and **EmbSpatial** (Du et al., 2024). Among them, VSI-Bench provides a fine-grained diagnosis over eight subcategories, including four numerical-answer tasks (*Object Count*, *Absolute Distance*, *Object Size*, *Room Size*) and four multiple-choice tasks (*Relative Distance*, *Relative Direction*, *Route Plan*, *Approaching Order*). To examine whether the proposed grounding mechanism preserves general multimodal ability, we further evaluate

Table 2: **Comparison on ReVSI (Zhang et al., 2026b).** Published baselines use their native frame settings. SpatialStack and GeoWeaver share a matched 32-frame protocol; GeoWeaver also uses 16 and all frames. Scores are accuracy (%). The 16-frame mean covers five subtasks, versus seven for 32 and all frames. Bold marks column maxima; parenthesized 32-frame gains are over SpatialStack-5B (52.0).

Methods	Frames	Avg.	Numerical Question				Multiple-Choice Question		
			Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
Cambrian-S-7B (Yang et al., 2025d)	128	49.1	**48.4**	60.5	65.5	46.7	37.1	48.5	37.0
VST-7B-SFT (Yang et al., 2025c)	4 FPS	46.4	35.4	52.6	67.9	47.2	49.2	36.9	35.4
SpaceR-7B (Ouyang et al., 2025a)	32	30.5	30.7	34.5	52.0	18.6	22.8	34.5	20.2
Spatial-MLLM-4B-135k (Wu et al., 2025a)	16	40.5	40.7	45.3	46.8	–	32.3	37.4	–
Spatial-MLLM-4B-820k (Wu et al., 2025a)	16	40.9	41.5	40.0	53.1	–	30.7	39.2	–
VLM-3R-7B (Fan et al., 2025)	32	50.1	41.6	61.6	64.8	52.5	46.5	49.5	34.1
SpatialStack-5B (Zhang et al., 2026a)	32	52.0	33.8	62.7	61.2	51.7	61.5	45.5	47.5
Ours									
GeoWeaver-5B	16	54.2	38.6	61.1	62.8	–	61.4	47.0	–
	32	54.9 (+2.9)	38.0	63.4	63.6	**54.2**	**65.3**	46.9	53.1
	All	54.1	38.6	61.3	62.2	53.7	63.8	46.1	53.1
GeoWeaver-10B	16	57.4	40.5	64.5	69.3	–	62.1	50.5	–
	32	**57.5 (+5.5)**	42.6	**66.8**	**69.8**	50.9	64.4	**52.4**	**55.4**
	All	57.3	43.2	66.5	69.7	50.7	65.0	51.7	54.7

Table 3: **Spatial ability evaluation.** Comparison with reported methods on six spatial reasoning benchmarks. Parenthesized gains are relative to the highest listed baseline in each column.

Methods	VSI-Bench	SPAR-Bench	ViewSpatial	BLINK	3DSRBench	EmbSpatial	Overall
VLM-3R-7B (Fan et al., 2025)	60.9	42.4	40.5	52.3	51.5	68.2	52.6
MindCube-3B (Yin et al., 2025)	17.2	20.8	24.1	35.1	2.8	37.0	22.8
SpatialLadder-3B (Li et al., 2025d)	44.9	32.9	39.9	43.0	42.8	58.2	43.6
Spatial-MLLM-4B (Wu et al., 2025a)	46.3	35.3	34.7	40.5	36.2	50.0	40.5
SpaceR-7B (Ouyang et al., 2025a)	41.5	34.2	35.9	49.6	40.5	66.9	44.8
ViLaSR-7B (Wu et al., 2025b)	44.6	37.4	35.7	51.4	46.6	67.3	47.2
Cambrian-S-7B (Yang et al., 2025d)	62.9	37.9	41.3	37.9	45.0	72.8	49.6
GeoWeaver-5B	**72.7** (+9.8)	**69.2** (+26.8)	**52.8** (+11.5)	**58.1** (+5.8)	**58.6** (+7.1)	**77.5** (+4.7)	**64.8** (+12.2)

on **MMBench** (Liu et al., 2024a), **Video-MME** (Fu et al., 2025), **BLINK**, and **TempCompass** (Liu et al., 2024b).

Metrics. For VSI-Bench, we report numerical MRA, multiple-choice accuracy, and the eight-category mean (Avg.) under EASI. Other benchmarks use their standard scores.

Implementation Details. We instantiate GeoWeaver on Qwen3.5 (Qwen Team, 2026) multimodal LLMs with 4B and 9B backbones, using the native vision encoder of each backbone to extract semantic visual tokens. For geometric grounding, we use a pretrained VGGT encoder as a frozen geometry branch. Each selected geometry layer is aligned to the visual-token resolution through a 2×2 spatial merge and projected into the LLM hidden space by a shared MLP. The grounding stage assigns compact geometric evidence to each visual token through token-conditioned layer selection, with two geometry layers selected by default. The selected evidence is injected into visual tokens through residual grounding, where the output projection is zero-initialized for stable fine-tuning. During training, the native vision encoder, visual merger, and VGGT remain frozen, while the grounding layers are fully optimized. We fully fine-tune the 4B language-model backbone in GeoWeaver-5B and train language-model LoRA adapters for the 9B backbone in GeoWeaver-10B.

4.2 Main Results

Results on VSI-Bench. Table 1 reports detailed results on VSI-Bench. GeoWeaver-5B achieves **72.7** Avg., outperforming SpatialStack-5B, the strongest listed open-source baseline, by **5.2** points. Scaling to GeoWeaver-10B further improves the score to **75.05**, yielding a **7.6**-point gain over that baseline. Relative to SpatialStack-5B, GeoWeaver-5B improves object counting, object size estimation, and route planning; GeoWeaver-10B also improves relative direction. These category scores indicate that the complete grounding pipeline provides spatial evidence for object-centric and navigation-oriented reasoning. GeoWeaver also scores above the listed general-purpose VLMs, including GPT-5, Gemini-3-Pro, and Qwen3-VL-8B. The allocation variants below separately test the contribution of token-specific depth selection within GeoWeaver.

Results on ReVSI. Table 2 evaluates GeoWeaver on ReVSI, a frame-aware spatial reasoning benchmark where the available visual evidence depends on the input frame budget. This setting stresses

Table 4: **General capability evaluation.** Scores on four multimodal benchmarks and their arithmetic mean. Parenthesized GeoWeaver differences are relative to SpatialStack-5B; individual benchmark changes can differ from the aggregate.

Method	MMBench	Video-MME	BLINK	TempCompass	Overall
Qwen3.5-4B	83.25	62.44	**61.12**	66.84	68.41
SpatialStack-5B (Zhang et al., 2026a)	**83.42**	63.74	55.46	69.37	68.00
GeoWeaver-5B	83.33	**64.80** (+1.06)	58.10 (+2.64)	**69.69** (+0.32)	**68.98** (+0.98)

the ability to infer geometric relations from partial temporal observations, especially for distance estimation, relative relations, and route planning. GeoWeaver-5B achieves 54.9 average accuracy with 32 frames, and GeoWeaver-10B reaches 57.5. In the matched 32-frame comparison, GeoWeaver-5B improves over SpatialStack-5B from 52.0 to 54.9 and scores higher on all seven subtasks (Table S2). Among the settings covering all seven subtasks, 32 frames give the highest GeoWeaver aggregate; increasing the budget further does not monotonically improve the result.

Results on spatial reasoning benchmarks. Table 3 compares GeoWeaver with representative methods for which complete six-benchmark rows are available. GeoWeaver leads the methods listed in the table on *every* benchmark and obtains an overall score of **64.8**, a **12.2**-point gain over the strongest complete listed baseline. The gain is especially large on **SPAR-Bench** (**69.2**, +26.8 relative to the highest listed baseline), which evaluates relational understanding and scene-level spatial structure. We report SpatialStack separately on VSI-Bench in Table 1 and in the matched-frame ReVSI comparison in Table S2; a single-protocol six-benchmark SpatialStack row is not available from the underlying records.

The gains span benchmarks emphasizing local spatial perception as well as compositional and viewpoint-aware reasoning. Together with the allocation ablation and cross-view correspondence probe, this pattern supports token-adaptive multi-level geometry as a useful source of spatial evidence across diverse settings.

General capability evaluation. An important question is how spatial grounding affects general multimodal ability. Table 4 reports a four-task mean of **68.98** for GeoWeaver, above both SpatialStack-5B (68.00) and the base Qwen3.5-4B (68.41). GeoWeaver improves Video-MME from 62.44 to **64.80** and TempCompass from 66.84 to **69.69** relative to the base model, with a similar MMBench score (83.33 vs. 83.25). BLINK is lower than the base model (58.10 vs. 61.12), although higher than the SpatialStack value reported in this table (55.46). Thus, the stronger spatial results coexist with a higher general-task mean and a benchmark-specific tradeoff.

Discussion. Across the evaluations, GeoWeaver combines strong spatial reasoning with a competitive general-task aggregate. The matched ReVSI comparison and cross-view correspondence probe further show gains in downstream reasoning and in the geometry-sensitive information carried by visual tokens. Within the proposed architecture, the allocation ablation identifies token-specific depth selection as an important contributor to these results.

Matched-frame and correspondence analyses. With 32 input frames for both systems, GeoWeaver-5B scores 54.90 versus 52.00 for SpatialStack-5B on ReVSI, with higher scores on all seven subtasks (Table S2). We also evaluate cross-view correspondence on 30 ScanNet scenes and 1,400 directed frame pairs using a common 12×16 candidate grid. Relative to the native visual tokens before geometry grounding, grounded tokens improve pooled R@1 from 20.75 to 24.34 and R@5 from 41.32 to 48.02 (Table S3). The paired scene-macro R@1 gain is 3.45 points, positive in all 30 scenes. This probe evaluates the complete grounding operation. Table 5 separately tests token-specific evidence allocation.

4.3 Ablations

Effect of geometry layer selection. Table 5 shows that geometry layer selection substantially affects grounding quality. Using the first 12 layers achieves 66.31 on VSI-Bench, while uniformly sampling 12 layers improves the score to 69.65. The latter 12 layers perform best with 72.7, indicating that deeper geometry features provide more reasoning-relevant structural abstractions. This result supports our motivation that geometry layers are heterogeneous and that spatio-temporal reasoning benefits from a geometry bank composed of more structured evidence.

Table 5: **Ablation study on VSI-Bench.** We analyze GeoWeaver from five aspects: geometry bank construction, geometry bank size, evidence compactness, evidence allocation granularity, and integration design.

Variant	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
A. Geometry Bank Construction									
First 12 Layers	66.31	81.37	50.70	**77.00**	54.89	64.50	73.38	50.00	78.67
Uniform 12 Layers	69.65	81.11	53.48	76.90	57.50	67.50	**85.01**	55.00	80.67
Latter 12 Layers	**72.7**	**83.20**	**61.80**	76.80	**62.40**	**72.80**	83.50	**56.30**	**84.70**
B. Geometry Bank Size									
4 Layers	68.45	80.05	55.22	75.15	58.02	67.05	79.72	51.15	81.25
8 Layers	70.45	81.45	58.55	76.05	60.25	69.35	81.65	53.45	82.85
12 Layers	**72.7**	**83.20**	**61.80**	**76.80**	**62.40**	**72.80**	**83.50**	**56.30**	**84.70**
16 Layers	71.95	82.65	60.85	76.45	61.65	72.15	82.75	55.05	84.05
All Layers	71.53	82.30	60.10	76.30	61.20	71.50	82.40	54.60	83.80
C. Evidence Compactness									
1 Selected Layer	71.69	81.94	60.25	76.37	59.77	**73.50**	83.27	53.75	84.67
2 Selected Layers	**72.7**	**83.20**	**61.80**	76.80	62.40	72.80	**83.50**	**56.30**	**84.70**
3 Selected Layers	72.54	83.03	61.35	76.68	63.41	73.00	83.18	55.00	84.67
4 Selected Layers	72.09	82.87	60.58	**76.92**	**64.09**	73.00	82.79	52.50	84.00
D. Evidence Allocation Granularity									
Uniform Averaging	70.44	82.10	58.40	76.30	60.10	69.20	82.10	52.50	82.80
Global Layer Weights	71.30	82.50	59.80	76.50	61.10	70.40	82.60	54.10	83.40
Token-Adaptive Allocation	**72.7**	**83.20**	**61.80**	**76.80**	**62.40**	**72.80**	**83.50**	**56.30**	**84.70**
E. Integration Design									
Input-Level Fusion	67.49	78.20	54.60	74.10	58.30	66.20	79.10	49.00	80.40
Decoder-Side Fusion	69.725	80.10	57.30	75.40	60.20	68.70	81.60	52.50	82.00
Pre-Reasoning Grounding	**72.7**	**83.20**	**61.80**	**76.80**	**62.40**	**72.80**	**83.50**	**56.30**	**84.70**

Effect of geometry bank size. We vary the size of the geometry bank in Table 5. Four, eight, and twelve layers score 68.45, 70.45, and 72.7, respectively; expanding to sixteen or all layers lowers the score to 71.95 and 71.53. The non-monotonic result favors the twelve-layer bank, which provides multi-level evidence without aggregating every encoder depth.

Effect of compact geometry grounding. We study how many geometry layers should be selected for each visual token. Selecting one layer obtains 71.69, while selecting two layers achieves the best score of 72.7. Increasing the number to three or four layers decreases the score to 72.54 and 72.09, respectively. These results suggest that a single layer lacks complementary evidence, whereas overly broad selection introduces redundant or weakly related cues. Thus, compact evidence selection provides a better balance between geometric diversity and noise suppression.

Effect of evidence allocation granularity. We compare evidence allocation within the same grounding architecture. Uniform averaging scores 70.44, and shared global layer weighting scores 71.30. Token-adaptive allocation reaches 72.7, gains of 2.26 and 1.40 points, respectively, supporting per-token selection from the multi-level geometry bank. **Effect of integration design.** We compare three implemented integration designs. Input-level fusion achieves 67.49, decoder-side fusion reaches 69.725, and the proposed pre-reasoning grounding design obtains 72.7. The proposed combination of integration site and token-adaptive allocation performs best in this comparison. Because these implementations also differ in how evidence is allocated, the result complements the allocation ablation rather than isolating the effect of placement alone.

5 Conclusion

GeoWeaver selects compact evidence for each visual token from a frozen geometry encoder's multi-level features before language decoding. It leads the listed methods across six spatial benchmarks, improves over SpatialStack in the matched 32-frame ReVSI setting, and raises the four-task general multimodal mean over the base model. The allocation ablation and cross-view probe support token-specific geometric evidence as a useful representation for spatial reasoning.

Limitations. We evaluate a frozen VGGT backbone with a fixed twelve-layer bank, leaving transfer to other geometry encoders and token grids untested. The bank uses hidden features rather than explicit depth maps, point maps, or camera poses; combining these structured states with token-adaptive selection remains for future study.

REPRODUCIBILITY STATEMENT

The method and training objective are described in Section 3. Implementation settings and additional benchmark results appear in Appendix A. Supplementary clarifications provide component-level training settings, a matched comparison with SpatialStack, cross-view correspondence analysis, and the inference-cost measurement protocol. Code and models will be released.

ETHICS STATEMENT

This study uses existing spatial reasoning datasets and introduces no new human-subject data collection. Any release of derived data will retain source attribution and follow the licenses and redistribution terms of the underlying datasets.

AI USE STATEMENT

During manuscript revision, we used generative AI tools to assist with literature checking, language editing, figure preparation, LaTeX formatting, and numerical consistency checks. Aggregate-score corrections were calculated from the reported category scores. This assistance did not produce new experimental measurements. The authors directed the revisions and take responsibility for the final content, including the scientific claims, reported results, citations, and figures.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*, 2025.

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.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025a.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Holistic evaluation of multimodal llms on spatial intelligence. *arXiv preprint arXiv:2508.13142*, 2025b.

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, 2024.

QwenLM Team (Alibaba Cloud). Qwen3-vl: Multimodal large language model series. https://github.com/QwenLM/Qwen3-VL, 2025. GitHub repository; accessed: 2025-11-14.

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.

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

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

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, 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*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Gemini. Gemini 3 Pro Model Card. Technical report, Gemini, November 2025. Accessed: 2025-11-18.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, Chenxin Li, Panwang Pan, Junbin Lu, Jingyan Jiang, Xinghao Ding, Yue Huang, and Zhi Wang. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world, 2026. URL https://arxiv.org/abs/2603.12746.

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

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3600–3610, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

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

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. *arXiv preprint arXiv:2505.21500*, 2025c.

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

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

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, 2025e. URL https://arxiv.org/abs/2510.08531.

Guoshan Liu, Bin Zhu, Yian Li, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. Enhancing action and ingredient modeling for semantically grounded recipe generation. *arXiv preprint arXiv:2602.15862*, 2026.

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

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025a.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *arXiv preprint arXiv:2509.18094*, 2025b.

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

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024b.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025c.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025d.

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

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*, 2024. URL https://arxiv.org/abs/2411.04923.

OpenAI. Gpt-5. https://openai.com/gpt-5/, 2025a. Accessed: 2025-11-11.

OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025b. Accessed: 2025-08-10.

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

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

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

ByteDance Seed Team. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025e.

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

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

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024.

xAI. Grok 4, 7 2025. URL https://x.ai/news/grok-4. Model announcement.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025c.

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. *arXiv preprint arXiv:2511.04670*, 2025d.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

Jiang Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. *arXiv preprint arXiv:2604.24300*, 2026b.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8995–9006, 2025b.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8600–8612, 2025.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025a.

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.

We provide additional details about the training and inference, as well as more experiments on various benchmarks and visualizations in Section A. Limitations are analyzed in Section B.

A Additional Details and Results

This appendix provides additional details for GeoWeaver. We first describe the implementation and training settings of the 5B and 10B variants. We then report additional ReVSI results under different frame budgets and provide qualitative visualizations that motivate token-adaptive geometric evidence allocation.

A.1 Implementation Details

Model configuration. GeoWeaver is instantiated with two model scales, denoted as GeoWeaver-5B and GeoWeaver-10B. They use Qwen3.5-4B and Qwen3.5-9B base multimodal models, respectively. In both variants, the native vision encoder and visual merger extract semantic visual tokens, while a pretrained VGGT supplies the geometry branch. All three components remain frozen throughout training. Following the main paper, we collect the latter half of the VGGT hidden states to construct the multi-level geometry bank and discard non-patch tokens before evidence allocation.

Grounding module. Each selected geometry layer is first normalized and spatially aligned to the visual-token resolution through a 2×2 spatial merge. A shared MLP then projects the aligned geometry tokens into the hidden space of the language model. During token-adaptive evidence allocation, each visual token predicts its layer preferences over the geometry bank and selects two geometry layers by default. The selected evidence is aggregated and injected into visual tokens through residual grounding. The output projection of the residual grounding branch is initialized to zero for stable optimization.

Training strategy. For GeoWeaver-5B, we fully finetune the language-model backbone and grounding module; the native vision encoder, visual merger, and VGGT stay frozen. For GeoWeaver-10B, we train language-model LoRA adapters and the full grounding module while keeping those same three components frozen. LoRA uses rank $r = 64$, scaling factor $\alpha = 128$, and dropout 0.05. The learning rate and global batch size are unchanged between settings. All experiments use 8 A100 GPUs (80GB) and the standard autoregressive next-token objective, without auxiliary reconstruction, contrastive, or geometry-specific supervision losses.

Table A: **Implementation summary of GeoWeaver.** The 5B variant fully finetunes the language model; the 10B variant uses language-model LoRA. Both fully train the grounding module and freeze the native vision encoder, visual merger, and VGGT.

Setting	GeoWeaver-5B	GeoWeaver-10B
Base multimodal model	Qwen3.5-4B	Qwen3.5-9B
Native vision encoder / merger	Frozen / frozen	Frozen / frozen
Geometry encoder	Frozen VGGT	Frozen VGGT
Geometry layers	Latter half	Latter half
Geometry alignment	2×2 spatial merge	2×2 spatial merge
Selected geometry layers	2	2
Frames	8	8
Language-model finetuning	Full-parameter	LoRA
Grounding-module finetuning	Full-parameter	Full-parameter
LoRA rank r	–	64
LoRA α	–	128
LoRA dropout	–	0.05
Learning rate	1e-5	1e-5
Global batch size	64	64
GPUs	8 A100	8 A100

Table B: **Frame-budget analysis on ReVSI.** All scores are reported as accuracy (%). The 16-frame setting does not include room size estimation or route planning. The best result in each column is highlighted in **bold**.

Methods	Frames	Avg.	Numerical Question				Multiple-Choice Question		
			Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
GeoWeaver-5B	16	54.2	38.6	61.1	62.8	–	61.4	47.0	–
	32	54.9	38.0	63.4	63.6	**54.2**	**65.3**	46.9	53.1
	64	53.6	38.7	62.3	62.8	52.1	61.6	45.7	51.6
	All	54.1	38.6	61.3	62.2	53.7	63.8	46.1	53.1
GeoWeaver-10B	16	57.4	40.5	64.5	69.3	–	62.1	50.5	–
	32	**57.5**	42.6	**66.8**	**69.8**	50.9	64.4	**52.4**	**55.4**
	64	56.9	42.4	65.7	69.6	50.7	65.0	51.8	53.0
	All	57.3	**43.2**	66.5	69.7	50.7	65.0	51.7	54.7

Table C: **Detailed results on SPAR-Bench.** We follow the original SPAR-Bench table format and report the fine-grained scores of GeoWeaver-5B. GeoWeaver-5B achieves the best overall score and shows strong gains on medium- and high-level spatial reasoning, especially position matching, distance inference, object spatial relation reasoning, and spatial imagination.

Methods	Rank	Avg.	Low	Low-level Spatial Perception							Medium	Medium-level Spatial Reasoning			High	High-level Spatial Reasoning							
				Depth-OC	Depth-OC-MV	Depth-OO	Depth-OO-MV	Dist-OC	Dist-OC-MV	Dist-OO	Dist-OO-MV		PosMatch	CamMotion	ViewChgI		Dist-OO	Dist-OO-MV	ObjRel-OC-MV	ObjRel-OO	ObjRel-OO-MV	Splmag-OC	Splmag-OC-MV	Splmag-OO	Splmag-OO-MV
GPT-4o (OpenAI, 2025b)	3	38.11	36.88	51.22	44.69	21.21	19.33	41.40	44.90	36.34	35.96	26.49	27.74	25.25	19.99	43.80	65.00	64.88	44.75	50.82	43.21	29.84	32.56	27.81	35.29
GPT-4.1	2	41.60	41.95	48.25	42.66	20.86	23.58	58.29	52.33	48.65	40.44	44.02	59.29	28.75	22.02	42.93	71.76	67.26	46.25	54.95	41.00	30.38	29.65	20.20	24.93
Doubao-1.5	1	41.74	33.24	34.89	37.22	25.24	21.40	35.04	34.99	40.58	36.57	47.36	55.73	39.00	29.47	49.49	74.71	69.64	35.75	70.33	47.65	34.95	33.14	35.76	42.86
SpaceR-7B (Ouyang et al., 2025a)	4	37.55	30.90	36.14	35.16	15.40	19.25	40.13	35.07	36.97	29.16	31.93	38.68	40.00	17.23	45.61	62.35	61.61	52.25	51.92	46.81	37.90	36.05	24.83	34.17
InternVL2.5-8B+ (Zhang et al., 2025)	–	63.25	**65.53**	81.53	**79.22**	**38.25**	**35.51**	**78.93**	**79.18**	**68.13**	**63.50**	63.01	78.88	**73.00**	**37.14**	61.29	86.47	79.76	64.00	69.23	59.00	47.31	50.00	42.38	53.50
GeoWeaver-5B	**1**	**69.15**	59.30	**82.94**	76.19	25.83	19.89	77.30	77.48	66.03	48.75	**75.13**	**86.26**	64.00	35.14	**80.36**	**89.12**	**83.93**	**91.25**	**85.44**	**83.38**	**73.39**	**72.09**	**66.23**	**78.43**

A.2 Additional Results

ReVSI evaluates spatial reasoning under frame-aware settings, where the available visual evidence changes with the input frame budget. This protocol is suitable for evaluating whether a model can recover geometric relations from partial temporal observations. Table B reports GeoWeaver results under different frame budgets.

Among the settings that cover all seven ReVSI subtasks, 32 frames gives the highest aggregate for both model scales. GeoWeaver-5B reaches 54.9 average accuracy, while GeoWeaver-10B reaches 57.5. Increasing the frame budget to 64 or all frames does not consistently improve performance in the evaluated configurations. The larger model configuration also scores higher at 32 frames.

SPAR-Bench. Table C reports the detailed SPAR-Bench results following the original benchmark format. GeoWeaver-5B achieves a 69.15 overall score, outperforming the strongest listed baseline. The improvement is most evident on medium- and high-level reasoning tasks, including position matching, distance inference, object spatial relation reasoning, and spatial imagination. These results indicate that the proposed grounding pipeline is effective when spatial reasoning requires object-level relations and structured scene understanding.

Dyn-Bench. Table D reports category-level results on Dyn-Bench. GeoWeaver-5B achieves the best overall score of 61.3 among the compared models. The strongest gain appears on Object-Scene reasoning, where GeoWeaver reaches 77.1 and outperforms all listed baselines. This category requires modeling how objects interact with scene structure under dynamic visual changes, which aligns with our design of grounding visual tokens with token-specific geometric evidence. The results suggest that pre-reasoning geometric grounding is particularly effective for scene-anchored spatial-temporal reasoning.

A.3 Additional Ablation Results

We provide detailed ablation results for geometry layer selection and evidence compactness. These results complement the main paper by showing how the construction and use of the geometry bank affect spatio-temporal reasoning.

Table F studies how many geometry layers should be selected for each visual token. Selecting two layers achieves the best overall score of 72.7, compared with 71.69, 72.54, and 72.09 when selecting one, three, and four layers. This pattern supports a compact evidence budget within token-adaptive allocation; the separate uniform and global aggregation controls are reported in Table 5.

Table D: **Category-level comparison on Dyn-Bench (Huang et al., 2026).** We aggregate the original Dyn-Bench subcategories into Inter-Object, Object-Scene, and Camera-Object groups following the benchmark (Huang et al., 2026). Best and second-best results are **bolded** and underlined, respectively.

Methods	Inter-Object	Object-Scene	Camera-Object	Avg.
Proprietary Models				
GPT-4o (Hurst et al., 2024)	46.5	63.7	46.1	50.1
GPT-5 (OpenAI, 2025a)	54.7	70.2	**59.2**	59.5
Gemini-2.5 Pro (Comanici et al., 2025)	56.1	64.4	55.8	59.8
Open-source General Models				
InternVL3-14B (Wang et al., 2025c)	54.0	71.6	43.2	53.7
InternVL3.5-8B (Wang et al., 2025d)	50.8	65.3	42.7	50.3
Qwen2.5-VL-7B (Bai et al., 2025b)	50.8	69.9	42.1	51.6
LLaVA-OneVision-1.5-4B (An et al., 2025)	49.8	64.6	40.9	49.9
LLaVA-OneVision-1.5-8B (An et al., 2025)	54.0	73.1	43.2	53.8
Spatial MLLMs				
SpaceR-7B (Ouyang et al., 2025b)	56.2	72.7	48.6	56.5
VST-7B-RL (Yang et al., 2025b)	56.3	74.4	45.7	55.7
Spatial-SSRL-7B (Liu et al., 2025d)	47.5	69.4	36.7	45.9
SpatialLadder-3B (Li et al., 2025e)	52.0	72.7	44.0	53.6
Region-level MLLMs				
UniPixel-3B (Liu et al., 2025b)	54.6	73.2	46.3	55.4
UniPixel-7B (Liu et al., 2025b)	**56.4**	75.4	50.1	58.1
VideoGLaMM (Munasinghe et al., 2024)	35.0	37.4	23.3	30.7
Sa2VA-InternVL2.5-8B (Yuan et al., 2025)	49.7	66.9	40.1	49.4
Sa2VA-InternVL3-14B (Yuan et al., 2025)	52.7	72.3	43.8	53.6
Sa2VA-Qwen2.5-VL-7B (Yuan et al., 2025; Bai et al., 2025b)	50.3	66.7	42.4	50.3
Sa2VA-Qwen3-VL-4B (Yuan et al., 2025; Cloud), 2025)	48.8	67.5	40.8	49.8
GeoWeaver-5B	54.6	**77.1**	52.3	**61.3**

Table E: **Ablation on geometry layer selection.** The latter 12 geometry layers provide the best overall performance.

Geometry Layers	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
First 12 Layers	66.31	81.37	50.70	77.00	54.89	64.50	73.38	50.00	78.67
Uniform 12 Layers	69.65	81.11	53.48	76.90	57.50	67.50	**85.01**	55.00	80.67
Latter 12 Layers	**72.7**	**83.20**	**61.80**	76.80	**62.40**	**72.80**	83.50	**56.30**	**84.70**

Table E studies how the construction of the geometry bank affects GeoWeaver. Using the first 12 VGGT layers yields 66.31, while uniformly sampling 12 layers improves the score to 69.65. The latter 12 layers achieve the best overall performance of 72.7, with clear gains on absolute distance, room size, relative distance, route planning, and appearance order. This trend shows that the selected depth range matters for downstream reasoning. The gap between uniform sampling and latter-layer selection further suggests that simply increasing layer diversity is not sufficient; the bank also needs features that are useful for spatio-temporal reasoning.

A.3.1 Evidence Compactness

Figure A illustrates evidence selection for two ROIs from the same input image. Their dense router probability distributions concentrate on a subset of geometry layers, and the selected layers differ between the ROIs. For one reference ROI, we also visualize grounded response maps under budgets of one, two, four, and all selected layers. The response patterns change with the budget, while the two-layer setting yields a focused spatial response in this example. The quantitative compactness ablation in Table 5 shows that selecting two layers gives the highest VSI-Bench score among the tested budgets.

A.3.2 Geometry Layer Selection

To inspect how GeoWeaver allocates geometric evidence, we visualize the spatial distribution of geometry-layer preferences in Figure B. For each spatial location, we take the top-k geometry layers according to the dense router scores, compute their average layer index, and project that value back to the image plane. The heatmap therefore represents the average index of the top candidates, not the probability or selection frequency of an individual layer. In the displayed examples, this index varies across image regions: some tokens favor shallower candidates and others deeper candidates. The

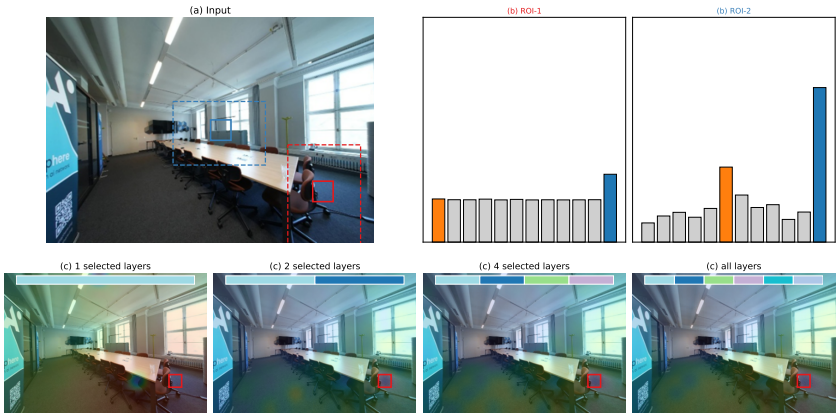


Figure A: **Visualization of evidence compactness.** Dense router distributions for two ROIs in one image, followed by grounded response maps for one reference ROI using one, two, four, or all geometry layers. The selected layers and response patterns vary with the ROI and evidence budget; the quantitative budget comparison is given in Table 5.

Table F: **Effect of evidence compactness.** We vary the number of geometry layers selected for each visual token during token-adaptive evidence allocation. Two selected layers achieve the highest overall score among the four tested budgets.

Selected Layers	Avg.	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
1 Layer	71.69	81.94	60.25	76.37	59.77	**73.50**	83.27	53.75	84.67
2 Layers	**72.7**	**83.20**	**61.80**	76.80	62.40	72.80	**83.50**	**56.30**	**84.70**
3 Layers	72.54	83.03	61.35	76.68	63.41	73.00	83.18	55.00	84.67
4 Layers	72.09	82.87	60.58	**76.92**	**64.09**	73.00	82.79	52.50	84.00

visualization illustrates the spatially varying decisions made by the token-conditioned router, while the allocation ablation measures their effect on benchmark performance.

Figure C visualizes how the same ROI interacts with geometry features from different VGGT layers. For each layer, we use the geometry token at the ROI location as a query and compute its cosine similarity with all tokens in that layer's feature map. The resulting heatmaps differ across layers. In the shown examples, L12 has a more localized response, L16 and L18 also respond to nearby structure, and L20 and L23 show broader responses. These layer-dependent patterns motivate a multi-level bank from which the router can choose evidence for each visual token.

B LIMITATIONS

GeoWeaver is designed as a modular grounding framework built on top of an external geometry encoder. This design allows the model to benefit from strong geometric priors without requiring additional 3D supervision, but it also means that further improvements may come from stronger or more domain-adaptive geometry encoders. In the current implementation, GeoWeaver uses hidden geometry features as the evidence bank, while explicit geometric states such as depth maps, point maps, or camera poses are not directly modeled. Incorporating these structured 3D representations may further benefit metric-scale reasoning and long-range spatial consistency.

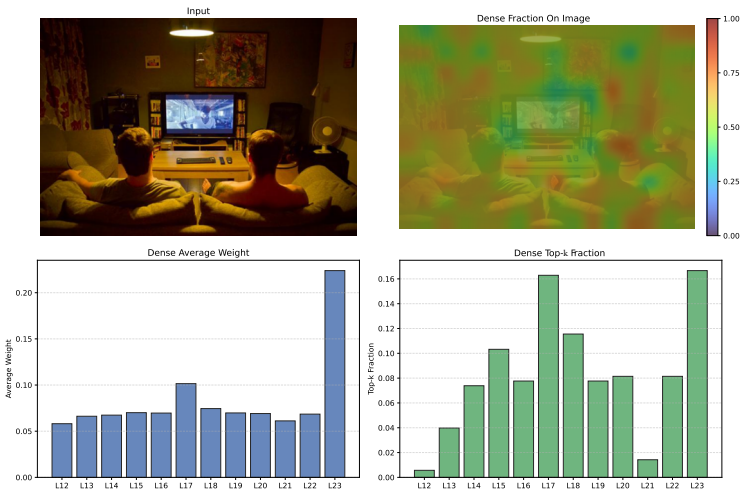


Figure B: **Visualization of geometry-layer preference.** For each spatial location, we compute the average layer index of the top-k candidates ranked by dense router scores and overlay it on the input image. Cooler and warmer colors indicate lower and higher average selected indices, respectively. The displayed examples illustrate spatial variation in layer preference.

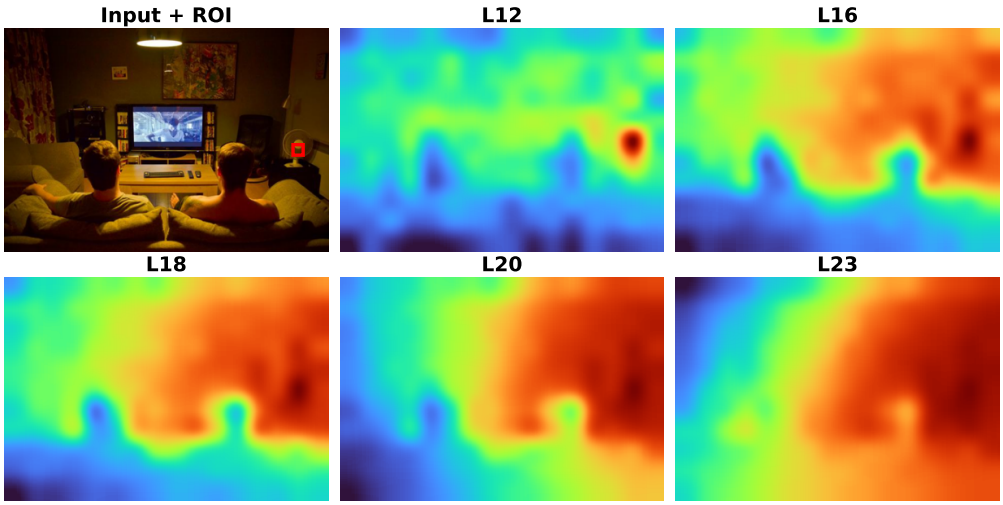


Figure C: **ROI-to-layer similarity visualization.** Given the same ROI, we use the geometry token at its location as a query and compute cosine similarity with all tokens in the same-layer feature map. The selected layers exhibit distinct response patterns in these examples, motivating a multi-level geometry bank for token-adaptive evidence allocation.

C Supplementary Results and Clarifications

C.1 Training Configuration and Model Size

The native Qwen3.5 vision encoder, its visual merger, and the VGGT geometry encoder are frozen in every reported GeoWeaver experiment. GeoWeaver-5B updates the language model and grounding module. GeoWeaver-10B updates language-model LoRA adapters and the grounding module. Thus, full-parameter fine-tuning in the 5B setting refers to the language model, while the visual and geometry encoders remain frozen. Table S1 lists the corresponding parameter counts. The 10B configuration trains approximately 116.39M language-model LoRA parameters and 75.59M grounding parameters, for a total of 191.98M trainable parameters.

Table S1: Model parameters and trainable components. B and M denote billions and millions of parameters. Total counts include frozen components.

Setting	GeoWeaver-5B	GeoWeaver-10B
Base multimodal model	Qwen3.5-4B	Qwen3.5-9B
Total parameters	5.50B	10.39B
Frozen VGGT parameters	0.909B	0.909B
Grounding module	55.92M	75.59M
Trainable parameters	≈4.26B	≈191.98M
Native vision encoder / merger	Frozen / frozen	Frozen / frozen
Language-model adaptation	Full tuning	LoRA
LoRA rank / scaling / dropout	–	64 / 128 / 0.05

Instruction tuning uses SPAR-234K, LLaVA-Hound-64K, VLM-3R ScanNet, and VSI appearance-order data, with sampling fractions of 60%, 60%, 60%, and 50%, respectively. Each percentage specifies the portion sampled from its own source; these percentages are not mixture weights that sum to one. The training configuration uses one epoch, AdamW with learning rate 10^{-5}, a global batch size of 64, cosine scheduling, warm-up ratio 0.03, weight decay 0.01, and seed 0. Training samples eight frames and uses eight A100-80GB GPUs. The reported results do not include variability across independently trained seeds.

The geometry bank comprises VGGT layers 12–23 using zero-based indexing, and each visual token selects two depths. Routing is conditioned on the native visual token. It selects geometry at the aligned spatial location across encoder depths; it does not search over all image positions. In the vectorized implementation, routing precedes the shared projector, which is applied only to the selected top-two geometry vectors. This is numerically equivalent to projecting all bank entries before top-two selection. The implementation retains the full frozen VGGT forward pass and adds no visual sequence tokens.

C.2 32-FRAME COMPARISON WITH SPATIALSTACK

SpatialStack-5B and GeoWeaver-5B share the Qwen3.5-4B native vision encoder and visual merger, together with the VGGT geometry encoder. Both systems freeze these encoders. Their geometry integration differs: SpatialStack injects multiple geometry levels into the language decoder, whereas GeoWeaver selects geometry depths separately for each visual token before language decoding.

Table S2 provides a matched 32-frame ReVSI comparison. GeoWeaver-5B improves the aggregate score by 2.90 points over SpatialStack-5B and improves each of the seven subtasks. GeoWeaver-10B obtains an aggregate score of 57.50. The 10B result additionally changes the base language-model scale, whereas the 5B comparison shares the base model and geometry backbone.

Table S2: ReVSI results with 32 input frames for every model. The aggregate follows the benchmark evaluation protocol.

Task	SpatialStack-5B	GeoWeaver-5B	GeoWeaver-10B
Object counting	33.81	38.00	42.60
Absolute distance	62.74	63.40	66.80
Object size	61.18	63.60	69.80
Room size	51.74	54.20	50.90
Relative distance	61.50	65.30	64.40
Relative direction	45.52	46.90	52.40
Route planning	47.52	53.10	55.40
Aggregate	52.00	54.90	57.50

For the frame-budget analysis, the 16-frame evaluation omits room size and route planning. Its aggregate therefore has a different task composition from the 32-frame, 64-frame, and all-frame

evaluations. Among the comparable seven-task settings, the 5B model scores 54.9, 53.6, and 54.1 with 32, 64, and all frames; the corresponding 10B scores are 57.5, 56.9, and 57.3. More input frames do not monotonically improve the reported score.

C.3 Cross-view correspondence analysis

Protocol. We evaluate correspondence retrieval on 30 ScanNet scenes from VSI-Bench, with eight selected frames per scene. Ground-truth correspondences use measured RGB-D data, camera intrinsics, and camera-to-world poses, retaining points that are visible in the target view according to its ground-truth depth. The resulting evaluation contains 1,400 directed frame pairs and 124,203 valid correspondences. All representations use a common 12×16 grid with 192 target candidates; raw VGGT features are 2×2 averaged onto this grid.

For each source descriptor, cosine similarity ranks target-view descriptors. R@1 and R@5 measure whether the ground-truth target index appears among the first one or five candidates. Mean reciprocal rank (MRR) averages the inverse rank of the target index. PCK@1 and PCK@2 measure whether the highest-scoring target lies within one or two grid units of the projected target location. Median error is measured in grid units. Chance retrieval on this grid is 0.52% for R@1 and 2.60% for R@5.

Native, geometry, and grounded representations. Table S3 reports the pooled correspondence results. Native Qwen3.5 tokens retain cross-view correspondence information, while every evaluated VGGT layer obtains higher R@1 and R@5. This diagnoses a difference in recoverable geometry under the descriptor-retrieval probe. Grounding improves the native-token R@1 from 20.75 to 24.34 and R@5 from 41.32 to 48.02, while reducing median error from 1.67 to 1.25 grid units. The native-token comparison with VGGT motivates geometry enrichment; the grounded-token comparison evaluates the resulting representation.

Table S3: Cross-view correspondence for native visual tokens, each depth of the geometry bank, and grounded tokens. Retrieval, MRR, and PCK values are percentages; median error is in grid units. VGGT depth indices are zero-based.

Representation	R@1	R@5	MRR	PCK@1	PCK@2	Med. err.
Qwen3.5 native	20.75	41.32	31.15	37.05	54.60	1.67
GeoWeaver grounded	24.34	48.02	36.07	43.34	64.47	1.25
VGGT layer 12	39.08	73.27	54.30	61.35	78.99	0.72
VGGT layer 13	40.26	74.82	55.58	63.98	82.46	0.69
VGGT layer 14	43.36	79.63	59.20	70.20	89.47	0.63
VGGT layer 15	40.64	77.20	56.67	66.49	86.78	0.67
VGGT layer 16	37.05	74.07	53.35	62.34	84.75	0.74
VGGT layer 17	39.76	78.13	56.48	67.38	90.08	0.68
VGGT layer 18	40.81	79.04	57.42	68.71	90.20	0.67
VGGT layer 19	33.47	70.27	49.85	57.60	81.54	0.82
VGGT layer 20	32.83	68.69	48.89	56.24	79.60	0.85
VGGT layer 21	32.23	68.53	48.42	56.21	80.42	0.85
VGGT layer 22	31.70	68.28	48.07	55.92	80.68	0.86
VGGT layer 23	43.15	83.14	60.34	73.12	93.05	0.63

The paired scene-macro R@1 improvement is 3.45 points, with a 95% bootstrap confidence interval of $[3.15, 3.77]$. All 30 scenes improve; a two-sided exact sign test gives $p = 1.863 \times 10^{-9}$. Scene-macro gains for R@5, PCK@1, and PCK@2 are 6.43, 6.03, and 9.51 points, respectively. These scene-level averages differ from differences between pooled scores because scenes contribute different numbers of correspondences. The uncertainty estimates characterize variation across the evaluated scenes, rather than across independent training runs. This probe validates the complete grounding operation; the allocation-rule ablation separately tests the token-adaptive routing mechanism.

Other geometry-augmented models. We apply the same scenes, frames, correspondence set, and candidate grid to VG-LLM and VLM-3R. Table S4 compares representations before and after

geometry integration within each model. VG-LLM's fixed additive pre-fusion improves scene-macro R@1 by 3.35 points, with 95% interval $[2.86, 3.86]$ and positive effects in all 30 scenes. VLM-3R's scene-macro effect is -0.07 before its projector, with interval $[-0.17, 0.01]$, and -0.07 after projection, with interval $[-0.20, 0.07]$. The latter intervals do not establish a positive correspondence gain under cosine retrieval. This result concerns patch identity under the descriptor probe, rather than the usefulness of VLM-3R's instruction-tuned fusion for downstream reasoning.

Table S4: Correspondence before and after geometry integration in other models. Scores are percentages except median error, which is in grid units.

Representation	R@1	R@5	PCK@1	PCK@2	Med. err.
VG-LLM: native	15.55	39.25	31.68	53.67	1.78
VG-LLM: additive fusion	19.12	47.93	38.92	65.87	1.32
VLM-3R: native, pre-projector	13.20	32.99	24.74	43.11	2.48
VLM-3R: fused, pre-projector	13.11	32.96	24.71	43.21	2.47
VLM-3R: native, projected	8.62	22.40	16.86	31.87	3.47
VLM-3R: fused, projected	8.52	22.32	16.82	31.79	3.47

SpatialStack integrates geometry inside the language decoder. We capture its visual-position hidden states after decoder layers 0, 1, and 2, and compare them with a forward pass through the same checkpoint with geometry disabled. Table S5 shows positive paired scene-macro R@1 effects after all three layers. At layer 2, PCK@1 and PCK@2 increase from 36.11 and 53.48 without geometry to 38.22 and 57.41 with geometry.

Table S5: SpatialStack correspondence at visual-token positions inside the decoder. Paired R@1 effects are scene-macro differences with 95% bootstrap intervals.

Capture point	No geometry R@1 / R@5	With geometry R@1 / R@5	Paired R@1 effect (95% interval)
After layer 0	19.75 / 39.42	20.17 / 40.19	$+0.46$ $[0.39, 0.54]$
After layer 1	19.80 / 39.42	20.77 / 41.47	$+0.96$ $[0.85, 1.07]$
After layer 2	20.20 / 40.28	21.18 / 42.21	$+0.93$ $[0.80, 1.06]$

Within-method paired effects keep the checkpoint and representation stage fixed while changing the geometry contribution. Across models, the visual backbones, projectors, feature dimensions, native grids, and training objectives differ. SpatialStack is additionally measured inside the decoder. Absolute scores therefore do not provide a controlled ranking of architecture families.

C.4 Inference efficiency

We measure model-only generation on one NVIDIA A800-SXM4-80GB GPU, using bfloat16 and batch size one. Eight 480×640 frames and the prompt yield 2,451 input tokens, and each run generates exactly 32 output tokens. CPU preprocessing and device transfer precede timing. Two warm-up runs are followed by five measured runs. Both systems use PyTorch 2.6.0+cu124 and the same Qwen3.5 PyTorch fallback for linear attention.

Table S6: Efficiency under a matched eight-frame, 32-output-token protocol. Latency is the mean and standard deviation over five runs.

Metric	SpatialStack-5B	GeoWeaver-5B
Generation latency (s)	1.822 ± 0.060	1.811 ± 0.040
Peak allocated memory (GiB)	15.505	15.505
Throughput (samples/s)	0.549	0.552
Output throughput (tokens/s)	17.56	17.67
Profiled computation (TFLOPs)	53.25	53.10

The latency differences in Table S6 lie within the observed runtime variation, supporting comparable inference cost for the two geometry-augmented systems under this protocol. Both include the cost of the frozen VGGT encoder. FLOPs are counted at the PyTorch operator level for the full 32-token generation; custom fused Triton or CUDA kernels may be omitted. These counts support a matched relative comparison rather than an analytical total.

C.5 Contribution scope and interpretation of the ablations

GeoWeaver belongs to the broad pre-fusion family: its grounded visual tokens enter the language model before decoding. Spatial-MLLM, VG-LLM, and VLM-3R also construct geometry-enhanced visual representations before language decoding. The specific mechanism studied by GeoWeaver is per-token sparse allocation across geometry-encoder depths. Each visual token selects its own top-two subset from a twelve-depth evidence bank, followed by residual grounding at the visual-language interface. The integration stage and evidence-allocation rule are separate design axes.

The allocation-granularity comparison obtains 70.44 with uniform averaging, 71.30 with globally learned layer weights, and 72.7 with per-token allocation. These results support token-specific selection over a shared layer preference. The comparison among input-level fusion (67.49), decoder-side fusion (69.725), and token-adaptive pre-reasoning grounding (72.7) evaluates complete geometry integration designs. Allocation changes along with the integration design, so the score differences do not isolate the causal effect of timing alone. The qualitative layer-response and routing visualizations illustrate spatially varying evidence selection, without assigning predetermined human semantic labels to individual geometry depths.

The correspondence probe measures geometry recoverable by descriptor retrieval, and the benchmark experiments measure downstream reasoning. Together, they support improved geometric representations and performance in the evaluated settings. Language reasoning, temporal modeling, supervision, and geometry-encoder quality remain complementary factors. In the general multimodal evaluation, GeoWeaver's aggregate of 68.98 is comparable to the base model's 68.41, while BLINK decreases from 61.12 to 58.10. The aggregate therefore does not imply preservation of every individual capability.

C.6 Qualitative comparison on VSI-Bench

Figure S1 contrasts GeoWeaver and SpatialStack on an appearance-order question. The models agree on the first two categories but differ in the order of the door and trash bin.

Question

What will be the first-time appearance order of the following categories in the video: lamp, chair, door, trash bin?

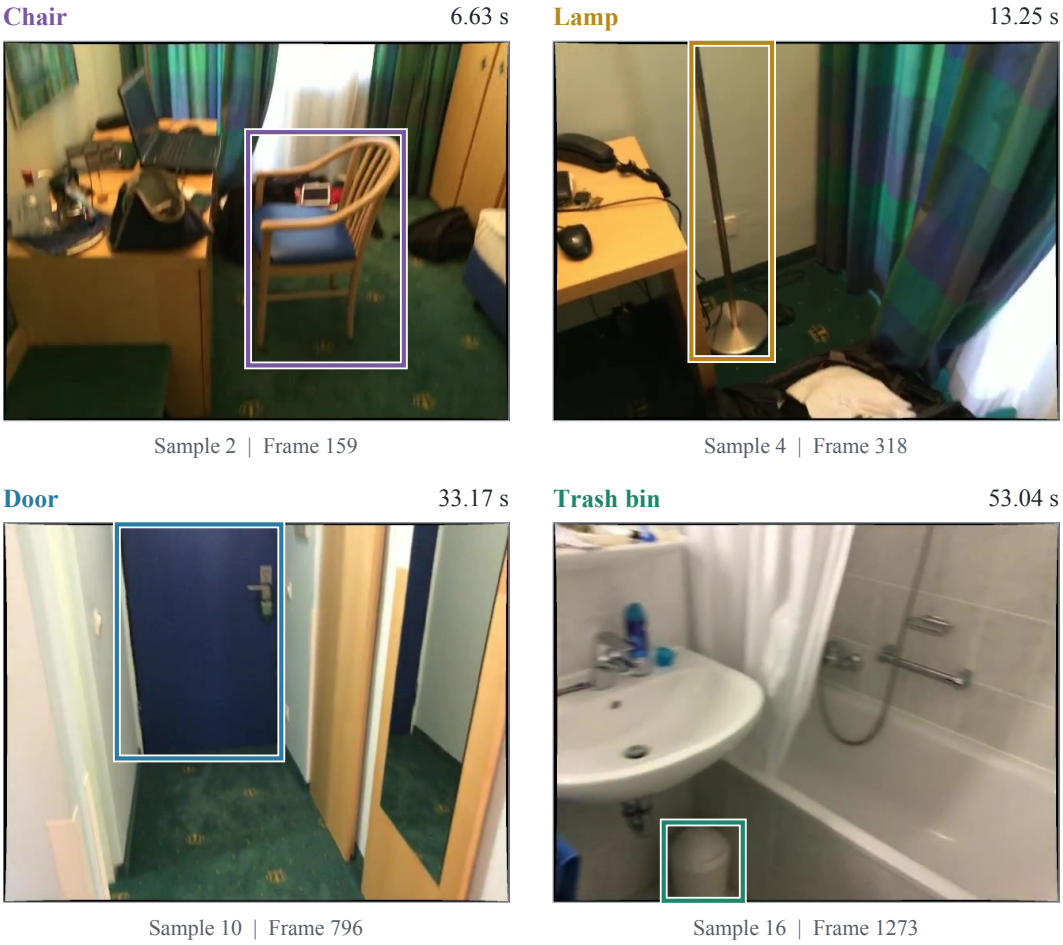


Figure S1: **Qualitative comparison on VSI-Bench.** Case 2239 from ScanNet scene0046_01. Four representative clear views from the 32-frame uniform input sampling grid are shown with approximate timestamps and manually added boxes; sample indices are zero-based. The first-appearance order is provided by the benchmark annotation. GeoWeaver matches this order, whereas SpatialStack reverses the door and trash bin.