GEOMETRY-AWARE MEMORY COORDINATION FOR SPATIAL UNDERSTANDING FROM LONG VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial understanding from long videos requires integrating partial observations across viewpoints and recalling past content within a limited memory budget. Repeated visits can reveal complementary views or changes in familiar regions. Existing memory methods use predictive novelty to guide retention or enrich language inputs with geometric features. Yet novelty under viewpoint changes need not reflect distinct content, and an observation can be redundant for geometry yet retain unique content for future questions. We introduce **GEOMEM**, a geometry-aware framework that coordinates visuospatial and episodic memories. Visuospatial memory provides historical context for geometry encoding of new views. The resulting features enrich the visual inputs used to form episodic memory, which retains observed content and its temporal context for later questions. Shared cross-view evidence guides separate retention decisions for geometric support and content under fixed budgets. Historical content can therefore remain accessible even when its source view becomes geometrically replaceable. Experiments on VSI-Bench and VSI-SUPER show leading performance in spatial understanding, long-horizon recall, and counting. Budget analyses show improved geometry estimation and question answering at matched memory budgets. Ablations support geometry fusion and complementary memory updates. Combining geometric association with content comparison outperforms either criterion alone. Streaming measurements show stable peak GPU memory and per-frame observation time on a four-hour video. Code and scripts will be released upon acceptance.

1 INTRODUCTION

Spatial understanding from long videos requires multimodal large language models (MLLMs) (Bai et al., 2025a) to infer relationships among objects and regions from observations collected across viewpoints (Yang et al., 2025). A moving camera sees only part of the environment at a time, and useful observations may be separated by long intervals. Repeated visits can reveal familiar regions from new viewpoints or capture changes in their contents. Answering later questions can therefore require relating the current view to previous observations and recalling content that is no longer visible (Yang et al., 2026b). Retaining all visual context, however, incurs growing computational and memory costs as videos lengthen (Zhang et al., 2024). This motivates selecting historical evidence that supports continued spatial understanding and future recall within a limited memory budget.

Existing approaches differ in how they represent and retain historical observations. **Visual Memory** methods (Yang et al., 2026b) use predictive surprise to guide compression and consolidation, supporting continued encoding and historical retrieval (Figure 1(a)). Under viewpoint changes, however, predictive novelty need not indicate additional geometric support or distinct historical content. **Geometry-enhanced Memory** methods (Liu et al., 2026b) fuse visual and geometric features, organizing recent and older observations into working and episodic memories (Figure 1(b)). Both memories enrich language inputs after geometry encoding, with feature similarity guiding episodic replacement. History can also support the geometric encoding of new observations. A view can be redundant for geometry yet still preserve unique content needed by future questions.

To support these complementary uses, we propose **GEOMEM**, a coordinated memory framework for spatial understanding from long videos. Drawing inspiration from how human memory connects spatial context with specific past experiences (Burgess et al., 2002; Tulving, 2002), we maintain

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.

Visuospatial Memory (VSM) and **Episodic Memory** (EM). VSM provides historical context for continued geometry encoding, while EM retains observed content and its temporal context for later questions. As shown in Figure 1(c), historical VSM context enriches the geometry features to form EM. Shared cross-view evidence then coordinates their updates through geometric coverage checks for VSM and content comparisons for EM. Separate retention decisions under fixed budgets allow an observation to remain in EM even when other retained views already provide its geometric support.

Specifically, visuospatial and episodic memories retain native key-value (KV) states in the geometry encoder (Zhuo et al., 2026) and language model (Bai et al., 2025a), respectively. For each frame, geometry features computed with VSM context are aligned and fused with visual features at the language model input. It reads recent episodic entries and produces visual KV for the current observation. VSM prunes redundant frames using geometric support from retained views. EM compares visual features at reliable correspondences, omitting repeated tokens only when their historical representatives remain in EM. Older candidates are discarded when necessary to meet the budgets. When a question arrives, the model retrieves relevant episodic entries to generate its answer.

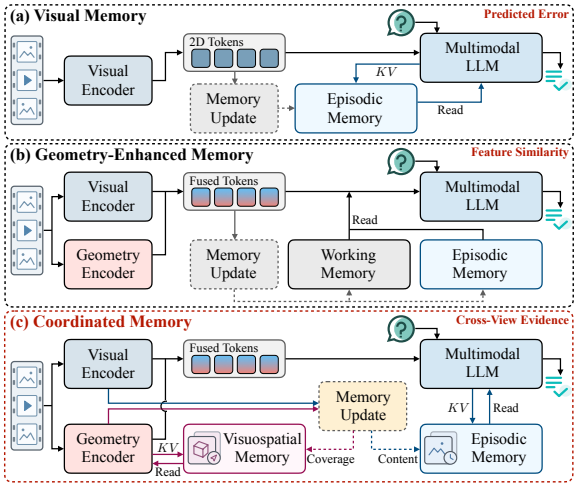


Figure 1: **Comparison of memory designs** for spatial understanding. **(a)** Visual memory uses prediction error to guide retention. **(b)** Geometry-enhanced memory uses feature similarity for episodic replacement. **(c)** GEOMEM uses shared cross-view evidence but separates geometric and content redundancy.

GEOMEM achieves leading results on VSI-Bench (Yang et al., 2025) for spatial understanding and VSI-SUPER (Yang et al., 2026b) for long-horizon recall and counting. Budget analyses show that selecting historical views by geometric support reduces depth and rotation errors, while episodic deduplication improves question answering at matched storage budgets. Ablations show complementary benefits from both update strategies. Geometry fusion further improves recall and counting when both updates are retained. Combining geometric association with content comparison outperforms either criterion alone. Streaming measurements show stable peak GPU memory and mean per-frame observation time on a four-hour video under fixed cache budgets. These findings support coordinating historical context for geometry encoding with content retention for future questions.

Our main contributions are as follows:

- **Coordinated Spatial Memory.** We introduce a dual-memory framework that separates history for continued geometry encoding from history retained for future recall. Coordinating visuospatial and episodic memories allows accumulated geometric context to strengthen the spatial representations of new views and enrich the evidence available for answering later questions.
- **Cross-View Memory Updates.** Shared cross-view evidence distinguishes geometric redundancy from content redundancy. Separate retention decisions allow distinct episodic content to remain accessible after its source frame becomes geometrically replaceable, reducing repeated storage under independent memory budgets.
- **Performance and Analyses.** GEOMEM achieves leading results on Benchmarks. Budget studies and ablations support geometry fusion and complementary memory updates. Streaming measurements show stable memory usage and per-frame observation time under fixed budgets.

2 RELATED WORK

Spatial Understanding from Videos. Recent MLLMs increasingly study spatial understanding directly from RGB images and videos, reducing reliance on explicit 3D inputs. Spatial-MLLM (Wu et al., 2025) augments visual representations with geometric features from a visual geometry foun-

dation model. Video-3D LLM (Zheng et al., 2025b) and Ross3D (Wang et al., 2025a) further exploit video observations and geometric cues for 3D scene understanding. Other methods improve spatial reasoning through geometry-aware representations or explicit reasoning mechanisms (Chen et al., 2026b; Zheng et al., 2025a; Huang et al., 2025; Lu et al., 2026; Fan et al., 2026). More recent approaches explore scene-centric modeling and adaptation during inference, including GPT4Scene (Qi et al., 2026), SmartMage (Zhang et al., 2026), and Spatial-TTT (Liu et al., 2026a). These methods primarily improve how spatial information is represented, reasoned over, or adapted. Our work instead focuses on how spatial and historical evidence should be retained during prolonged observation, where past views support both continued geometry estimation and later recall.

Memory for Long Video Understanding. Long video understanding requires compressing visual history while preserving spatial and temporal evidence for future queries. MovieChat (Song et al., 2024) consolidates redundant tokens through short- and long-term memory for long-context reasoning. Flash-VStream (Zhang et al., 2024) maintains compact memories for temporal context and visual detail. StreamChat (Xiong et al., 2025) uses hierarchical memory for streaming video reasoning and multi-turn interaction. Cambrian-S (Yang et al., 2026b) guides compression and consolidation with predictive surprise for long-horizon recall. VLM^2 (Liu et al., 2026b) incorporates geometry into working and episodic memories, using feature similarity for episodic replacement. Our framework coordinates geometric support for encoding and distinct content retention for future recall through shared cross-view evidence, with separate retention criteria under fixed budgets.

3 METHOD

3.1 GEOMETRY-AWARE MEMORY ENCODING

Overall Architecture. As shown in Figure 2, GEOMEM maintains two key-value (KV) caches: visuospatial memory (VSM) and episodic memory (EM). At step t, the geometry encoder reads VSM $\mathcal{M}_{\text{VSM}}^{(t-1)}$ to encode the RGB frame. The multimodal LLM (MLLM) reads recent entries in EM $\mathcal{M}_{\text{EM}}^{(t-1)}$ and forms visual KV from visual and geometry features fused at its input by the connector. Geometry predictions and visual token features provide cross-view evidence for updating both caches under fixed budgets. Given a question $\boldsymbol{x}_{\text{txt}}$, the MLLM answers using selected EM entries.

Visuospatial Memory Encoding. Following StreamVGGT (Zhuo et al., 2026), the geometry encoder processes one sampled frame $\boldsymbol{I}_t$ at each step t. At each cross-frame attention layer ℓ, queries from the current frame attend to its current keys and values together with the retained KV in $\mathcal{M}_{\text{VSM}}^{(t-1)}$. The geometry features $\boldsymbol{Z}_{\text{geo}}^{(t)}$ are the final geometry decoder outputs at current image token positions after interaction with historical VSM, excluding camera and register tokens. These features are passed to the connector. The encoder also predicts camera parameters, depth, and confidence for cross-view association. The current keys $\boldsymbol{K}_{\text{VSM}}^{(t,\ell)}$ and values $\boldsymbol{V}_{\text{VSM}}^{(t,\ell)}$ form native KV candidates grouped by source frame, with retention determined in Section 3.3.

Episodic Memory Encoding. The visual encoder independently encodes each sampled frame $\boldsymbol{I}_t$ into visual features $\boldsymbol{Z}_{\text{vis}}^{(t)}$ after spatial merging. We bilinearly resample $\boldsymbol{Z}_{\text{geo}}^{(t)}$ onto this visual token grid using the resize, crop, and padding mappings of both branches. Following the fusion operation in Spatial-MLLM (Wu et al., 2025), the connector projects both feature sets with MLPs and adds them to form the language model's visual input tokens. The MLLM reads recent entries in $\mathcal{M}_{\text{EM}}^{(t-1)}$ for historical observation context. At every decoder layer ℓ, hidden states derived from these fused inputs produce native keys $\boldsymbol{K}_{\text{EM}}^{(t,\ell)}$ and values $\boldsymbol{V}_{\text{EM}}^{(t,\ell)}$ at current visual positions as EM candidates.

3.2 CROSS-VIEW EVIDENCE

Geometric Association. We establish candidate token correspondences through projective data association (Newcombe et al., 2011). For frame i, we use predicted depth $\hat{\boldsymbol{D}}_i$, confidence $\hat{\boldsymbol{C}}_i$, intrinsics $\hat{\boldsymbol{A}}_i$, and pose $\hat{\boldsymbol{T}}_i$, which maps camera coordinates to a shared reference frame. Using the preprocessing mappings, we express current visual token n's center as $\boldsymbol{u}_t$ in geometry prediction coordinates. We back-project it to the 3D point $\boldsymbol{p}_t$ in the current camera frame and reproject it to $\boldsymbol{u}_i$

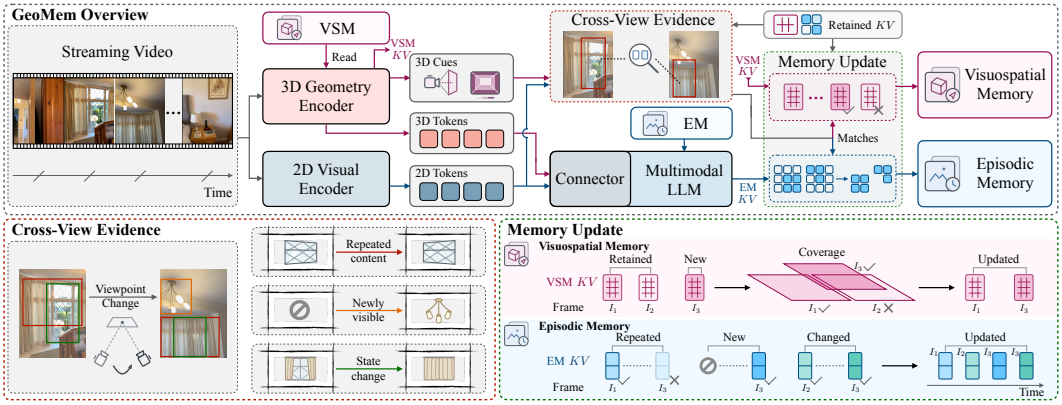


Figure 2: **Overview of GeoMem.** The geometry encoder reads visuospatial memory (VSM), while the multimodal LLM forms episodic memory (EM) from visual and geometry features fused by the connector. Cross-view evidence guides updates to both KV caches. VSM retains complete frame KV blocks for geometric support, while EM selects KV entries at visual token positions to reduce redundancy and preserve new evidence and historical states.

in historical frame i:

$$\begin{aligned} \boldsymbol{p}_t &= \hat{\boldsymbol{D}}_t(\boldsymbol{u}_t)\hat{\boldsymbol{A}}_t^{-1}\bar{\boldsymbol{u}}_t, \\ \boldsymbol{u}_i &= \pi\Big(\hat{\boldsymbol{A}}_i\big((\hat{\boldsymbol{T}}_i^{-1}\hat{\boldsymbol{T}}_t)\cdot\boldsymbol{p}_t\big)\Big), \end{aligned} \tag{1}$$

where $\bar{\boldsymbol{u}}_t$ is the homogeneous pixel coordinate, $\cdot$ denotes a rigid transformation acting on a 3D point, and π performs perspective division. A projection is valid when both positions lie in valid image regions with sufficient confidence and its transformed depth is positive and consistent with $\hat{\boldsymbol{D}}_i(\boldsymbol{u}_i)$. We map $\boldsymbol{u}_i$ back to the historical visual grid and require it to fall within the valid image region. The corresponding token m yields the candidate pair $((t,n),(i,m))$ in the correspondence set $\mathcal{C}$.

Content Comparison. For each pair $((t,n),(i,m)) \in \mathcal{C}$, we compare the full visual token features using cosine distance d_{vis}:

$$d_{\text{vis}} = 1 - \cos\Big(\boldsymbol{Z}_{\text{vis}}^{(t)}[n],\ \boldsymbol{Z}_{\text{vis}}^{(i)}[m]\Big), \tag{2}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity and $[n]$ indexes a token on the visual grid after spatial merging. Small distances identify redundancy candidates, while dissimilar pairs do not support redundancy removal. Changed or newly occluding surfaces may lack valid correspondences. All unmatched tokens bypass cross-view deduplication and follow the base retention policy.

3.3 Memory Update

We combine retained and current KV in the candidate cache $\widetilde{\mathcal{M}}_s^{(t)}$ and select indices $\mathcal{S}_s^{(t)}$:

$$\mathcal{M}_s^{(t)} = \widetilde{\mathcal{M}}_s^{(t)}[\mathcal{S}_s^{(t)}], \quad |\mathcal{S}_s^{(t)}| \le B_s, \quad s \in \{\text{VSM}, \text{EM}\}, \tag{3}$$

where $\mathcal{S}_s^{(t)}$ selects complete frame blocks for VSM or source token positions for EM, under budget B_s. Both caches may remain below budget. Selections are shared across each memory's layers, heads, and K/V pairs, while memories select independently and layers retain their own KV. Metadata include geometry predictions, visual token features for comparison, source frame indices, timestamps, and original positions. Each item is released only when neither memory references it.

Visuospatial Memory Update. We initialize $\mathcal{S}_{\text{VSM}}^{(t)}$ with the first reference frame (Xu & Oishi, 2026) and a recent frame window (Zhuo et al., 2026), both counted within B_{VSM}. We examine other frames newest first, projecting valid geometric samples into already selected views using the preceding checks. Following ORB-SLAM's keyframe redundancy principle (Mur-Artal et al., 2015), we skip a frame as redundant only when reliable geometric evidence is sufficient for evaluation and a sufficient fraction of its valid samples each receive support from the required number of selected views. Otherwise, we admit it while capacity remains, discarding older candidates at budget limit.

Episodic Memory Update. For each current token, we consider only historical correspondences to source token indices retained in $\mathcal{M}_{\text{EM}}^{(t-1)}$ and select the most recent reliable match. We tentatively omit the current token only when this match has sufficiently small d_{vis}, using the matched token as its representative. Tokens with different content or no reliable match enter as ordinary candidates without overwriting history. To enforce B_{EM}, we use the first-in, first-out (FIFO) option in Cambrian-S (Yang et al., 2026b), evicting the oldest entries by source time. If a representative is evicted, all current tokens it suppressed return as ordinary candidates subject to FIFO. Eviction and restoration continue until B_{EM} is met and every token omitted as redundant has a retained representative.

3.4 Training and Inference

Training Objective. Given a question $\boldsymbol{x}_{\text{txt}}$ and the EM KV $\mathcal{M}_{\text{read}}$ selected for answering, we supervise the reference answer $\boldsymbol{y}_{1:R}$ with token cross-entropy:

$$\mathcal{L}_{\text{QA}} = -\frac{1}{R}\sum_{r=1}^{R} \log p_{\boldsymbol{\theta}}(y_r \mid \boldsymbol{y}_{<r},\, \boldsymbol{x}_{\text{txt}},\, \mathcal{M}_{\text{read}})\,, \tag{4}$$

where R is the answer length and $\boldsymbol{\theta}$ denotes the model parameters. Following StreamVGGT (Zhuo et al., 2026), we distill camera parameters, depth, and point maps from a frozen bidirectional VGGT teacher to the causal geometry student. Both use the same training sequence, first-frame reference, and scale conventions. The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{QA}} + \lambda_g \mathcal{L}_{\text{GEO}}, \quad \mathcal{L}_{\text{GEO}} = \mathcal{L}_{\text{CAMERA}} + \mathcal{L}_{\text{DEPTH}} + \mathcal{L}_{\text{PMAP}}, \tag{5}$$

where λ_g weights geometry supervision. The camera term uses a Huber loss, while the depth and point map terms include confidence weighting and spatial gradient penalties.

Training Procedure. We jointly adapt pretrained models in one stage, training the connector, language model, and geometry decoder and heads. The MLLM visual encoder, geometry image backbone, and VGGT teacher remain frozen. Each finite video QA sample follows inference cache operations chronologically from empty caches, retrieving with questions alone. Retained and retrieved KV preserve gradients to the connector and earlier observation encoding, with discrete association and selection. Parameter updates follow complete sample or batch backpropagation, with caches reset between samples. Activation memory depends on sample length despite bounded KV caches.

Inference Procedure. Inference uses fixed weights and the causal geometry student, updating memories independently of questions. At question prefill, decoder layers independently read EM (Yang et al., 2026b) from shared storage indices. Frame scores average pairwise cosine similarities between each layer's question queries and each frame's retained visual keys. We read each frame's remaining KV together by descending score within B_{read} tokens per layer, skipping groups that do not fit. These per-layer read caches form $\mathcal{M}_{\text{read}}$ and remain fixed during generation with their original positions and timestamps. Question and answer positions continue from the end of the uncompressed observation sequence, identically across layers regardless of read cache lengths. New questions reselect $\mathcal{M}_{\text{read}}$. Question and answer KV are cached separately and never written to EM.

4 Experiments

4.1 Experimental Setup

Backbones, Datasets and Metrics. We instantiate GeoMem with Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-2B/8B (Bai et al., 2025a) as MLLM backbones. Following Zhuo et al. (2026), we initialize the geometry encoder from pretrained StreamVGGT and use a frozen VGGT teacher (Wang et al., 2025b) for geometric supervision on coherent video clips. Following Yang et al. (2026b), we perform single-stage joint training on approximately 1.3M samples: VSI-590K, 590K general video instruction samples from Cambrian-S-3M, and 120K image instruction samples from Cambrian-7M (Tong et al., 2024). Using RGB-only inputs, we report all eight subtask scores and their average on VSI-Bench (Yang et al., 2025), and VSR accuracy and VSC MRA across video durations on VSI-SUPER (Yang et al., 2026b).

Baselines. We compare GeoMem with proprietary multimodal models (GPT-5, Gemini-3-Pro, Gemini-2.5-Pro/Flash (Comanici et al., 2025), and Kimi-K2.5 (Team et al., 2026)), open-source

Table 1: **Evaluation on VSI-Bench.** GEOMEM achieves the highest average score among the methods. Backbones: [1] LLaVA-Video, [2] Qwen2-VL, [3] Qwen2.5, [4] Qwen2.5-VL, [5] Qwen3-VL.

Methods	Year	Average	Numerical Question				Multiple-Choice Question			
			Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models										
Gemini-2.5-Pro	2025	53.5	46.0	37.3	68.7	54.3	61.9	43.9	47.4	68.7
Gemini-3-Pro	2025	55.1	49.0	42.8	71.5	41.8	56.6	57.5	61.9	60.0
GPT-5	2025	55.0	53.3	34.4	73.3	47.5	63.7	48.6	50.2	68.9
Grok-4	2025	47.9	37.1	32.9	60.8	45.4	53.1	39.6	47.4	66.8
Kimi-K2.5	2026	54.5	57.2	34.9	69.3	54.4	59.6	41.3	52.1	67.0
Open-source General Models										
Qwen2.5-VL-3B	2025b	29.0	24.3	24.7	31.7	22.6	38.3	42.6	26.3	21.2
Qwen2.5-VL-7B	2025b	31.4	40.9	14.8	43.4	10.7	38.6	40.1	33.0	29.8
Qwen3-VL-2B	2025a	50.3	62.1	40.2	71.4	49.7	52.2	42.0	30.4	54.5
Qwen3-VL-8B	2025a	57.9	67.5	47.0	76.3	61.9	58.0	50.9	35.0	66.3
InternVL3-8B	2025	42.1	66.0	34.8	43.6	47.5	48.0	39.3	26.2	31.3
Spatial Intelligence Models										
Ross3D-7B[1]	2025a	22.8	20.5	–	–	–	–	35.1	27.3	–
Spatial-MLLM-4B[4]	2025	47.0	65.3	34.8	63.1	45.1	41.3	46.9	33.5	46.3
SpaceR-7B	2025	45.6	–	–	–	–	–	–	–	–
VLM-3R-7B[1]	2026	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
3DThinker-3B[4]	2026b	59.1	–	–	–	–	–	–	–	–
3DThinker-7B[4]	2026b	63.7	–	–	–	–	–	–	–	–
GS-Reasoner-7B[1]	2026a	64.7	69.1	61.9	70.0	65.7	65.4	88.9	44.3	52.3
VST-3B[4]	2026a	57.7	66.6	45.0	72.8	60.9	59.9	47.6	40.7	68.3
VST-7B[4]	2026a	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
VLM-IE3D-4B[4]	2026	47.6	67.5	38.5	55.0	58.7	47.7	45.1	36.6	31.9
Stream3D-VLM-4B[4]	2026	55.2	68.6	38.4	65.9	56.4	49.9	54.3	42.3	66.0
FactoSR-8B[5]	2026c	61.5	66.4	46.7	75.7	63.7	60.3	–	46.9	65.7
GPT4Scene-7B[2]	2026	25.6	38.1	–	–	–	–	38.7	28.8	–
SmartMage-8B[5]	2026	27.8	26.5	–	–	–	–	45.8	29.4	–
VLM^2-7B[1]	2026b	68.8	72.5	59.6	70.8	69.9	69.0	87.8	52.6	68.3
Spatial-TTT-2B[5]	2026a	64.4	70.8	47.8	71.7	65.9	61.8	73.0	47.4	77.0
Cambrian-S-3B	2026b	57.3	70.7	40.6	68.0	46.3	64.8	61.9	27.3	78.8
Cambrian-S-7B[3]	2026b	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
GEOMEM-2B[5]	2026	66.4	72.6	49.1	73.5	67.0	64.8	75.6	47.9	80.9
GEOMEM-7B[4]	2026	69.9	75.2	53.5	77.2	72.3	70.4	79.8	47.9	82.8
GEOMEM-8B[5]	2026	72.7	77.5	60.8	79.6	74.1	72.7	82.4	50.0	84.5

general-purpose MLLMs (Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a)), and spatial intelligence models (Video-3D LLM (Zheng et al., 2025b), Ross3D (Wang et al., 2025a), SpaceR (Ouyang et al., 2025), VLM-3R (Fan et al., 2026), 3DThinker (Chen et al., 2026b), GS-Reasoner (Chen et al., 2026a), GPT4Scene (Qi et al., 2026), SmartMage (Zhang et al., 2026), Spatial-TTT (Liu et al., 2026a), and Cambrian-S (Yang et al., 2026b)). For long-horizon recall and counting on VSI-SUPER, we additionally include MovieChat (Song et al., 2024) and Flash-VStream (Zhang et al., 2024). For both methods, we report the adaptations to Cambrian-S-7B-LFP provided in the Cambrian-S paper (Yang et al., 2026b). On this benchmark, Cambrian-S denotes its framework built on the LFP-finetuned model.

Implementation Details. We use 16 H800 80 GB GPUs, VERL (Sheng et al., 2025), FSDP, bfloat16, batch size 32. AdamW uses weight decay 0.01, clipping norm 1.0, and cosine decay after 3% linear warmup. Learning rates (LLM, connector, geometry decoder/heads) are $(2, 10, 1) \times 10^{-6}$. Geometry/appearance resolutions follow Zhuo et al. (2026); Wu et al. (2025), respectively. Training and evaluation VSI-Bench use 128 uniform frames and VSI-SUPER uses full videos at 1 FPS. Budgets are $B_{\text{VSM}} = 16$ frames, $B_{\text{EM}} = 16{,}384$ tokens, and $B_{\text{read}} = 8{,}192$ tokens/decoder layer. VSM protects the first frame and latest four. Observations read the four latest retained EM frame groups. Association uses frame-wise median confidence and 5% relative depth tolerance. Following Mur-Artal et al. (2015), VSM redundancy requires $\geq 90\%$ valid samples supported by three distinct selected views. Deduplication uses $d_{\text{vis}} \leq 0.05$. We set $\lambda_g = 0.1$ empirically and check early weighted gradient norms on the geometry decoder before clipping.

4.2 OVERALL PERFORMANCE

Spatial Understanding. Table 1 reports results on VSI-Bench. The 8B and 7B variants of GEOMEM achieve the highest and second-highest average scores, respectively. The 8B variant exceeds

Table 2: **Evaluation on VSI-SUPER.** Scores denote VSR accuracy and VSC MRA. GEOMEM achieves the best results across video durations on both tasks. OOC stands for out of context. Backbones: [1] LLaVA-Video, [2] Qwen2-VL, [3] Qwen2.5, [4] Qwen2.5-VL, [5] Qwen3-VL.

Method	Year	VSR (min)					VSC (min)			
		10	30	60	120	240	10	30	60	120
Qwen2-VL-7B	2024	26.7	28.3	28.3	–	28.3	–	–	–	–
Gemini-2.5-Flash	2025	–	–	41.5	OOC	–	–	–	10.9	OOC
Qwen3-VL-2B	2025a	35.0	30.0	28.3	0.0	–	0.8	0.0	0.0	0.0
MovieChat-7B[3]	2024	18.3	21.7	16.7	26.7	25.6	0.0	0.0	0.0	0.0
Flash-VStream-7B[3]	2024	28.3	33.3	23.3	28.3	31.7	0.0	0.0	0.0	0.0
Video-3D LLM-7B[1]	2025b	26.9	26.9	25.0	–	25.0	–	–	–	–
Ross3D-7B[1]	2025a	26.6	23.3	23.3	–	25.0	–	–	–	–
GPT4Scene-7B[2]	2026	26.9	26.6	28.3	–	26.6	–	–	–	–
SmartMage-8B[5]	2026	31.8	30.0	30.0	–	28.3	–	–	–	–
Spatial-TTT-2B[5]	2026a	38.3	35.0	28.3	30.0	–	31.8	45.6	36.2	38.4
Cambrian-S-7B[3]	2026b	45.0	41.7	40.0	40.0	40.0	40.6	42.0	35.0	34.0
GEOMEM-2B[5]	2026	43.3	43.3	41.7	41.7	41.7	41.0	41.6	36.6	33.8
GEOMEM-7B[4]	2026	48.3	45.0	45.0	43.3	43.3	43.2	46.0	37.8	38.8
GEOMEM-8B[5]	2026	53.3	51.7	50.0	50.0	48.3	48.2	53.0	45.2	47.0

the strongest baseline, VLM^2-7B, by 3.9 points and leads five of eight subtasks. The 2B variant also outperforms Spatial-TTT-2B across all eight subtasks with the same MLLM backbone. These results support our geometry-aware memory design across model sizes and question types. Strong performance on size and distance questions is consistent with using historical geometric context to enrich current visual representations. EM complements this spatial context by retaining observed content and its temporal information, providing evidence for questions about previously seen objects and their appearance order.

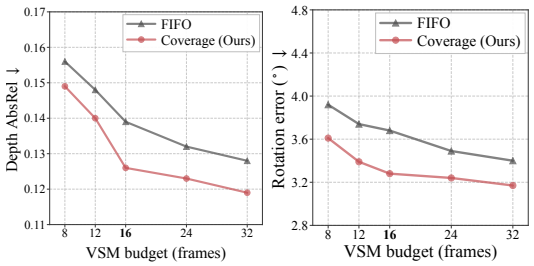
Long-Horizon Recall and Counting. Table 2 reports results on VSI-SUPER. The 8B and 7B variants of GEOMEM rank first and second, respectively, across all evaluated durations on both tasks. The 2B variant also outperforms Cambrian-S-7B on VSR from 30 to 240 minutes, demonstrating competitive recall with a smaller language backbone. These results support our coordinated memory framework for prolonged spatial observation. VSM provides historical context for current geometry encoding, enriching the representations written to EM. Separate retention decisions allow past content to remain available for later questions even when its source frames no longer provide unique geometric support. Cross-view comparison further reduces duplicate EM entries when retained representatives are available, leaving more capacity for distinct observed content.

4.3 IN-DEPTH ANALYSES

Visuospatial Memory Budget. Figure 3 compares Coverage with FIFO across VSM budgets from 8 to 32 frames. Both policies protect the first and four most recent frames. Starting from empty caches, we process 128 sampled frames per sequence using only the geometry branch and evaluate predictions from frame 33 onward with consistent scale handling. Both depth and rotation errors decrease as the budget grows, with Coverage outperforming FIFO throughout. Coverage with 16 frames even achieves lower errors than FIFO with 32 frames. These results support selecting historical context by geometric contribution. Removing redundant frames leaves capacity for complementary views that support subsequent geometry estimation beyond the protected recent window.

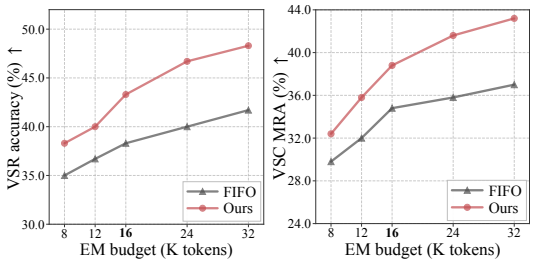
Episodic Memory Budget. Figure 4 compares our EM update strategy with FIFO on 120-minute videos, varying storage from 8K to 32K tokens (1K $=$ 1024 tokens). VSM uses Coverage with 16 frames, and reading is capped at 8K tokens per layer. Both VSR accuracy and VSC MRA improve with larger budgets, while our strategy consistently outperforms FIFO with a generally widening advantage. By reusing retained representatives for repeated content, our strategy leaves more capacity for distinct observations. Larger storage expands the historical evidence available to question-conditioned retrieval, supporting recall and counting under the same reading limit.

Streaming Inference Cost. Figure 5 compares our updates with dual FIFO under the default memory budgets. We process the same 240-minute video at 1 FPS on a single H800 80GB GPU with



(a) Depth AbsRel. (b) Rotation error.

Figure 3: **VSM budget and geometry estimation.** Both policies retain the first and four most recent frames. Coverage lowers depth and rotation errors. Bold ticks mark the default budget.



(a) VSR accuracy. (b) VSC MRA.

Figure 4: **EM budget and question answering.** On 120-minute videos, we fix VSM at 16 frames with Coverage and limit reading to 8K tokens per layer. Bold ticks mark the default budget.

Table 3: **Ablation of memory update strategies.** We fix the GEOMEM-7B checkpoint, encoders, geometry fusion, memory budgets, and reading policy. Disabled strategies use FIFO. VSR and regular VSC use 120-minute videos.

VSM Coverage	EM Dedup.	VSI-Bench Avg.↑	VSR-120 Acc.↑	VSC-120 MRA↑
×	×	64.6	33.3	30.8
✓	×	67.6	38.3	34.8
×	✓	66.7	40.0	35.8
✓	✓	69.9	43.3	38.8

Table 4: **Cross-view evidence for EM deduplication.** VSM Coverage remains enabled. Dedup. is the percentage of input visual tokens omitted as redundant after restoration, excluding FIFO evictions. Using both criteria performs best.

Criterion	VSR-120 Acc.↑	VSC-120 MRA↑	Dedup. (%)
FIFO	38.3	34.8	0.0
Feature only	40.0	36.2	36.8
Geometry only	38.3	35.4	43.1
Geometry + Content	43.3	38.8	28.6

batch size 1. Measurements cover frame encoding, cross-view association, and memory updates, excluding video I/O, decoding, and question answering. Both peak memory and mean per-frame observation time remain stable across the measured durations. Our updates require an additional 0.5–0.6 GiB and 36–41 ms per frame, reflecting the overhead of cross-view association and memory selection. This stability is consistent with fixed cache budgets limiting both the encoding context and the historical evidence examined during updates.

4.4 ABLATION STUDY

Memory Update Strategies. Table 3 evaluates both update strategies with the same checkpoint, budgets, and reading policy. Both encoders, geometry fusion, and caches remain active, with FIFO replacing disabled strategies. Each strategy improves all three metrics. Compared with deduplication alone, Coverage alone yields a larger gain on VSI-Bench but smaller gains on VSR and VSC. This pattern is consistent with their complementary roles. Coverage retains historical views that support geometry encoding, enriching representations used to form episodic entries. Deduplication reuses retained representatives for repeated content, leaving more capacity for distinct observations and temporal context. Combining both achieves best across three metrics, including a 10.0-point VSR gain over dual FIFO. These results support coordinating geometric support for continued encoding with content retention for future questions.

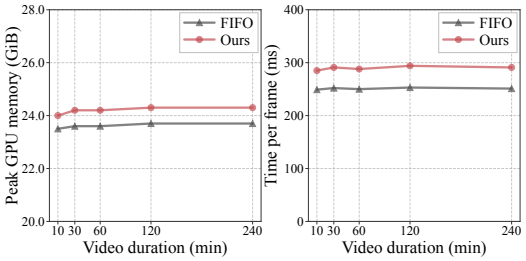


Figure 5: **Streaming observation cost** over 240 minutes with 16 VSM frames and 16K EM tokens. **Left:** cumulative peak GPU memory. **Right:** mean per-frame time over the preceding minute. Costs remain stable above dual FIFO.

Cross-View Evidence for Deduplication. Table 4 varies the EM redundancy criterion with VSM Coverage, budgets, and reading policy fixed. The reported ratio counts only redundant omissions

remaining after representative restoration, excluding FIFO evictions. Feature similarity alone improves both tasks over FIFO, whereas geometry alone leaves VSR unchanged despite the highest deduplication ratio. Combining geometry and content achieves the best results on both tasks while omitting 28.6% of input visual tokens, compared with 43.1% for geometry alone. Visual similarity can conflate observations from different locations, while spatial correspondence does not establish content redundancy. Using both criteria makes removal more selective, helping preserve distinct observations for future questions.

Geometry Fusion and Memory Updates. Figure 6 shows the full design outperforms both ablations on both tasks across all durations. The separately trained variant without fusion performs worse despite retaining both updates. This supports incorporating geometry into episodic representations in addition to using it for retention. Keeping fusion but replacing both updates with FIFO reduces performance. The VSR gap increases from 5.0 points at 10 minutes to 10.0 points at 120 minutes, suggesting that selective retention becomes more valuable as history accumulates. These results support our coordinated design, where geometric context enriches current observations and cross-view evidence guides the retention of geometric support and distinct historical content under fixed budgets.

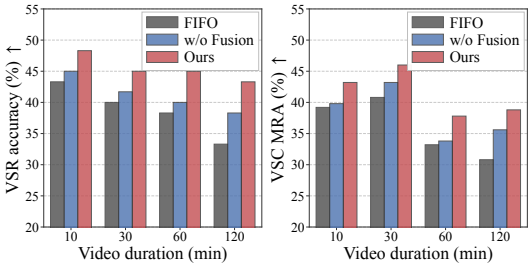


Figure 6: **Ablation of geometry fusion and memory updates.** FIFO retains fusion but replaces both memory updates using the Ours checkpoint. The separately trained w/o Fusion variant retains both updates. Under identical budgets, GEOMEM-7B achieves the best VSR and VSC results on VSI-SUPER across all durations.

Memory Reading Strategies. Figure 7 compares reading strategies using identical EM caches with a 16K token budget. All strategies read whole frame groups within the per-layer reading budget. Question-conditioned reading (Q-cond.) outperforms Recent and Uniform at restricted budgets. This suggests that temporal proximity or uniform coverage alone does not reliably identify the historical evidence needed by a specific question. The three strategies converge at 16K, when all retained entries can be read. Increasing the question-conditioned reading budget from 8K to 16K improves VSR and VSC by only 1.7 and 0.8 points, respectively. The default 8K budget therefore approaches the performance of reading the entire cache while limiting the context used for answering.

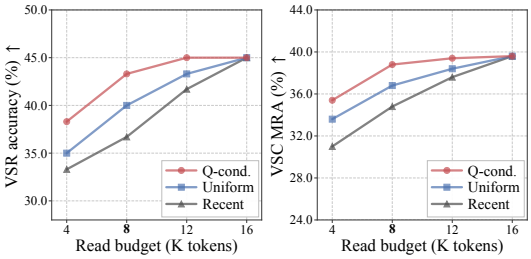


Figure 7: **Ablation of memory reading strategies.** Strategies share the same EM cache with a 16K token budget on 120-minute videos. Q-cond. (Question-conditioned) reading performs best under limited budgets. Bold ticks mark the default 8K reading budget.

5 CONCLUSION

We presented GEOMEM to address the different demands that continued geometry encoding and future questions place on visual history. The framework connects historical geometric context with the formation of new episodic representations, while distinguishing geometric support from content redundancy during memory updates. A view that becomes replaceable for geometry encoding can still contribute distinct content to later recall. Separate retention decisions allow the two memories to serve these different roles within fixed budgets. Evaluations on VSI-Bench and VSI-SUPER demonstrate strong spatial understanding, long-horizon recall, and counting across model sizes. Budget studies show the value of selecting geometric context and reducing repeated episodic storage. Ablations support geometry fusion and complementary memory updates, with joint geometric and content evidence improving question answering. Measured GPU memory usage and observation time remain stable on a four-hour video. These findings highlight the value of coordinating how history shapes new representations with how observed content is retained for future questions.

LLM USAGE STATEMENT

LLMs were used to assist with code organization, experimental consistency checks, and language polishing. All research ideas, experiments, results, and analyses were developed and verified by the authors, who take full responsibility for the paper.

REFERENCES

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, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Neil Burgess, Eleanor A Maguire, and John O'Keefe. The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641, 2002.

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2624, 2026b.

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.

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.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3d-aware representation supervision for scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wenhao Li, Xueying Jiang, Quanhao Qian, Deli Zhao, Ran Xu, Shijian Lu, and Gongjie Zhang. 3d-aware vlms with implicit and explicit geometries. In *European Conference on Computer Vision*, pp. 168–185. Springer, 2026.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-ttt: Streaming visual-based spatial intelligence with test-time training. In *European Conference on Computer Vision*, pp. 339–357. Springer, 2026a.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-language memory for spatial reasoning. In *European Conference on Computer Vision*, pp. 297–316. Springer, 2026b.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Haoqian Kang, Manyuan Zhang, Yan Feng, Ke Chen, and Yaowei Wang. Affin-space: Learning affine-invariant representations for 3d spatial understanding with mllms. In *International Conference on Machine Learning (ICML)*, 2026.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pp. 127–136. Ieee, 2011.

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

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 157277–157302, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

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

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9275–9286. IEEE, 2025a.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations*, 2025.

Zhisong Xu and Takeshi Oishi. Framevggt: Coherence-preserving memory for bounded streaming geometry. *arXiv preprint arXiv:2603.07690*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

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

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, 2026b.

Yijun Yang, Shenghe Zheng, Wenbo Li, Jianhui Liu, Haoze Sun, Yanbing Zhang, Jiaxiu Jiang, Lin Song, Haoyang Huang, Nan Duan, et al. Unfold the world: Factorize 4d properties in reinforcing spatial reasoning. In *European Conference on Computer Vision*, pp. 605–622. Springer, 2026c.

Hanxun Yu, Xuan Qu, Lei Ke, Boqiang Zhang, Yuxin Wang, Jianke Zhu, and Dong Yu. Stream3d-vlm: Online 3d spatial understanding with incremental geometry priors. In *European Conference on Computer Vision*, pp. 130–148. Springer, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

Yue Zhang, Yingzhao Jian, Yunqiu Xu, Xiaoxiao Sun, and Hehe Fan. Smartmage: Dynamic modality orchestration for 3d scene understanding. *arXiv preprint arXiv:2608.05137*, 2026.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing MLLMs with 3d vision geometry priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8995–9006. IEEE, 2025b.

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*, 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.