Q-GEOMEM: QUESTION-GUIDED GEOMETRIC MEMORY FOR VIDEO SPATIAL REASONING

Anonymous authors
Paper under double-blind review

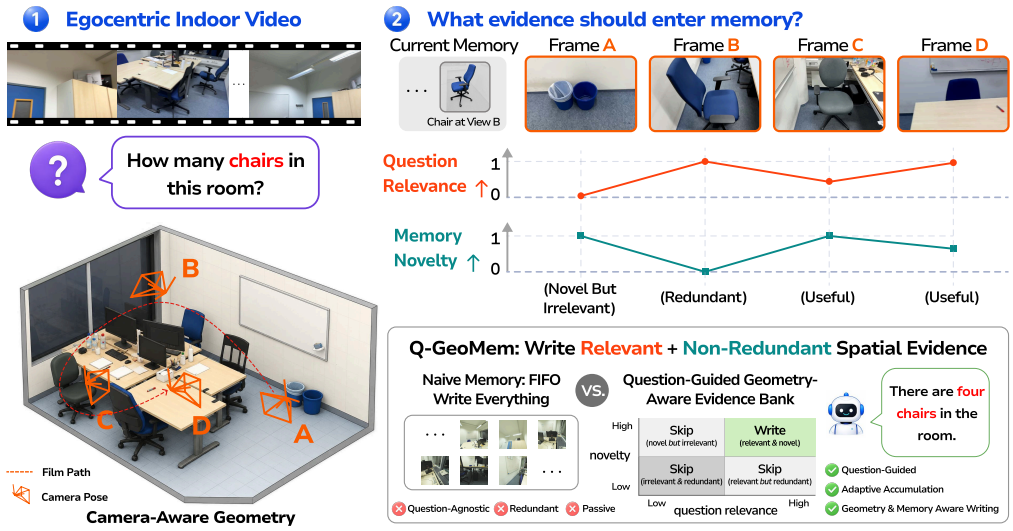


Figure 1: **Motivation of Q-GeoMem.** Egocentric videos reveal spatial layout through partial, camera-dependent views. For a question such as "How many chairs are in this room?", a chronological memory may mix useful chair observations with irrelevant or repeated views, whereas Q-GeoMem retains question-relevant, non-redundant, and geometry-grounded evidence.

ABSTRACT

Video spatial reasoning requires accumulating viewpoint-dependent evidence over time while retaining information useful to the question being asked. Existing spatial video-language models improve geometric perception and long-range context modeling, but often treat memory as a generic temporal cache, which can introduce redundant or irrelevant evidence and weaken long-horizon reasoning. We propose **Q-GeoMem**, a question-guided geometric memory framework for video spatial reasoning. Q-GeoMem injects camera-conditioned geometry into visual tokens and maintains a Fine-Grained Context Bank for recent dense context and a Semantic-Geometric Evidence Bank for compact long-range evidence. Each candidate frame is scored by calibrated question relevance and novelty with respect to the active bank, and the resulting utility drives both memory replacement and reading. Experiments on two in-domain and five out-of-distribution benchmarks show that Q-GeoMem achieves state-of-the-art performance in the evaluated settings, and controlled analyses validate question-guided geometric evidence selection.

1 INTRODUCTION

Memory turns observation into understanding. Perception alone offers only fleeting fragments of the world, while understanding and reasoning emerge from retaining what has been seen, integrating it with new evidence and preserving a coherent representation of the scene. This requirement is especially important in video spatial reasoning, where object relations, camera motion, distances, and directions are revealed gradually from viewpoint-dependent visual evidence.

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.

Recent progress in video spatial reasoning has been driven by stronger benchmarks and increasingly spatially grounded video-language models. Benchmarks such as Yang et al. (2025a;c), together with spatial-temporal evaluation suites such as Li et al. (2025b); Jia et al. (2026), show that video-based spatial intelligence requires more than short-range perception: models must preserve viewpoint-dependent observations, integrate geometric cues over time, and reason about order, distance, and direction from partial visual evidence. In response, recent methods have advanced along three directions: geometry-grounded perception and fusion (Zheng et al., 2025; Fan et al., 2025; Zhao et al., 2025; Li et al., 2026), spatial reasoning supervision and curricula (Ouyang et al., 2025; Li et al., 2025a), and memory-augmented temporal reasoning (Liu et al., 2025; Qian et al., 2024; Xiong et al., 2025; Zhang et al., 2025; Qian et al., 2025). Together, these methods improve how spatial evidence is perceived, trained, and retained.

Despite this progress, a central question remains open: **what spatial evidence should be remembered, in what representation, and under what update criterion**. Geometry-enhanced methods mainly improve how evidence is formed: VLM-3R (Fan et al., 2025) grounds video reasoning in instruction-aligned 3D reconstruction, and SpaceMind (Zhao et al., 2025) shows that camera-guided fusion is more effective than naive multimodal fusion. Yet they leave open how evidence should be retained for a specific question. Memory-oriented methods address temporal retention more directly. VLM^2 (Liu et al., 2025) demonstrates that persistent memory is valuable for long-horizon spatial reasoning, yet prior memory-augmented designs often use memory primarily as a repository of generic history rather than explicitly deciding, before writing, whether a candidate observation is relevant to the question and complementary to existing evidence. A similar gap appears in broader long-video methods (Qian et al., 2024; Xiong et al., 2025; Zhang et al., 2025; Qian et al., 2025), which maintain history but are not designed around camera-conditioned spatial evidence. As a result, current systems can retain more temporal context without retaining the evidence most useful for the question.

We argue that the missing piece is the coupling between question relevance, geometric reliability, and memory update. As illustrated in Fig. 1, memory should act as a question-guided evidence manager rather than a passive chronological cache: camera-conditioned geometry improves the reliability of each observation, the question determines which observations are useful, and novelty with respect to existing memory discourages redundant storage. The key challenge is thus not to remember more video, but to construct a compact memory that retains sufficient evidence for answering the question.

Motivated by this perspective, we propose **Q-GeoMem**, a memory-centric architecture for video spatial reasoning. Q-GeoMem injects camera-conditioned geometry into visual tokens and maintains two complementary memory banks: the Fine-Grained Context Bank (FGCB) keeps recent dense features and camera states for local, view-dependent details, while the Semantic-Geometric Evidence Bank (SGEB) stores compact long-range evidence scored by calibrated Q-Former-based question relevance and bank-relative novelty. This utility is recomputed against the active bank to bias memory reading and drive capacity-based replacement, and both banks are read before update and adaptively fused into the current feature.

Our contributions are three-fold:

- We introduce a **memory-centric formulation** of video spatial reasoning, in which camera-conditioned geometry, question relevance, and novelty jointly determine what evidence should be retained.
- We propose **Q-GeoMem**, a unified architecture that combines camera-guided geometry fusion, a Fine-Grained Context Bank for recent dense evidence, a Semantic-Geometric Evidence Bank for compact long-range evidence, and adaptive fusion of memory readouts.
- Extensive experiments and ablations show that selecting memory entries by both question relevance and novelty yields a more effective memory mechanism for video spatial reasoning.

2 Related Work

MLLMs for Video Understanding and Reasoning. Recent video MLLMs improve long-horizon reasoning by selecting and organizing temporally distributed evidence before answering, e.g., through question- or task-aware frame selection (Buch et al., 2025; Hu et al., 2025; Tang et al., 2025) or hierarchical querying at multiple temporal granularities (Azad et al., 2025). These methods

target general video understanding: they do not model how question relevance should interact with viewpoint-dependent spatial evidence, nor how such evidence should be retained once the relevant frame has passed.

Geometry-Augmented Video Spatial Reasoning. VG-LLM (Zheng et al., 2025) injects video-derived 3D geometry priors into visual tokens, and VLM-3R (Fan et al., 2025) couples geometry-aware modeling with reconstructive instruction data and introduces VSTI-Bench. Subsequent work studies how geometry should be supervised, fused, or selected: SpaceR (Ouyang et al., 2025) and SpatialLadder (Li et al., 2025a) use reinforcement-style training and progressive curricula, SpaceMind (Zhao et al., 2025) proposes camera-guided modality fusion, and GeoThinker (Li et al., 2026) advocates active rather than indiscriminate geometry integration. Related embodied studies ground spatial or affordance-relevant cues and apply deliberative reasoning to spatial decision making (Gao et al., 2025; Song et al., 2025). These works improve how spatial evidence is formed; we instead ask which camera-calibrated observations should be retained over time for the question being asked.

Memory-Augmented Video Understanding. Long-video memory methods either maintain explicit stores of selected observations updated online (Xiong et al., 2025; Huang et al., 2025; Zhang et al., 2025; Diko et al., 2025) or propagate history as a bounded latent recurrent state (Wang et al., 2025b; Ren et al., 2025; Shen et al., 2026); neither is specialized for camera-conditioned spatial evidence. The closest line to our work designs memory specifically for spatial reasoning over video. VLM^2 (Liu et al., 2025) introduces a dual-memory architecture with working memory for recent context and episodic memory for long-horizon recall. However, its memory update is primarily driven by learned saliency and similarity-based replacement, and the retained memory is not explicitly selected according to the question being asked. In contrast, Q-GeoMem scores each entry by calibrated question relevance and bank-relative novelty, and uses this utility for both read-time attention bias and capacity-based replacement.

3 METHOD

3.1 OVERVIEW

Given a monocular video and a question, Q-GeoMem builds a question-guided, geometry-aware memory for video spatial reasoning (Fig. 2). It first calibrates each frame's visual tokens with camera and geometry cues (Sec. 3.2), then maintains a Fine-Grained Context Bank (FGCB) for recent dense features and a Semantic-Geometric Evidence Bank (SGEB) for compact question-relevant evidence (Sec. 3.3), and finally fuses the retrieved evidence with the current observation for answer generation (Sec. 3.4).

3.2 CAMERA-GUIDED GEOMETRY FUSION

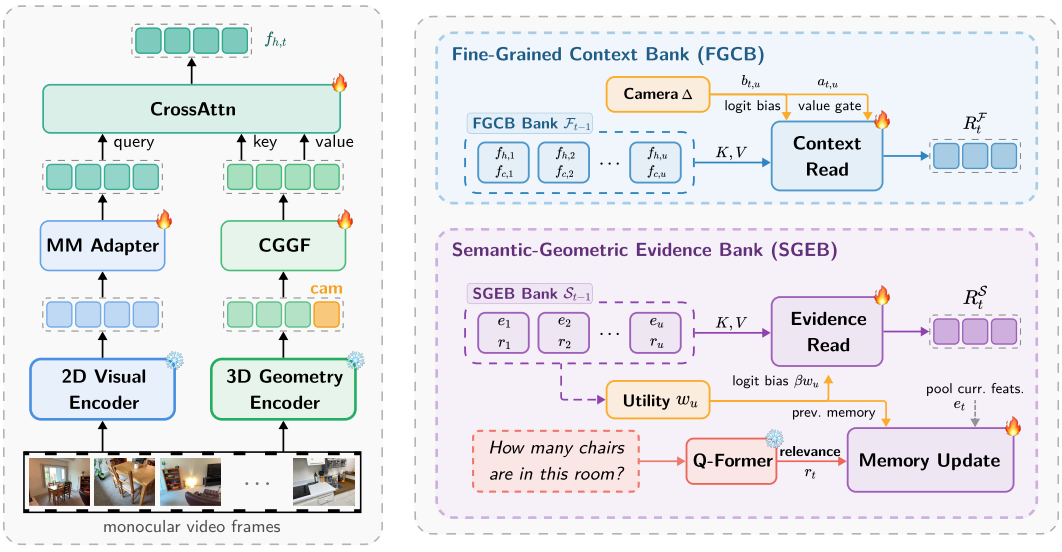
Before memory construction, we fuse visual tokens with camera-conditioned geometry, following the camera-guided fusion of SpaceMind (Zhao et al., 2025) so that both FGCB and SGEB operate on spatially grounded features. The motivation is that the same geometric observation may support different spatial relations under different camera states.

For frame t, let $f_{v,t} \in \mathbb{R}^{N_v \times d}$, $f_{g,t} \in \mathbb{R}^{N_g \times d_g}$, and $f_{c,t} \in \mathbb{R}^{1 \times d_g}$ denote the visual, geometry, and camera tokens, where $f_{c,t}$ is broadcast along N_g when needed. We condition the geometry on the camera token and predict a geometry-only reliability gate:

$$\mathbf{b}_g = \mathrm{MLP}_b\big([W_g^g f_{g,t};\, W_c^g f_{c,t}]\big), \quad \mathbf{r}_g = \sigma\big(\mathrm{MLP}_r(W_g^g f_{g,t})\big), \tag{1}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes channel-wise concatenation (with the frame-level camera feature broadcast along N_g), $\mathbf{b}_g \in \mathbb{R}^{N_g \times d}$, $\mathbf{r}_g \in \mathbb{R}^{N_g \times 1}$, and projection matrices with superscript g are learnable projections for geometry fusion. The visual stream then attends to the camera-conditioned geometry representation to obtain a geometry residual:

$$\hat{f}_{h,t} = \mathrm{Attn}\big(W_q^g f_{v,t}, W_k^g f_{g,t} + \mathbf{b}_g, \big(W_v^g f_{g,t} + \mathbf{b}_g\big) \odot \mathbf{r}_g\big)\,, \tag{2}$$



(a) Camera-Guided Geometry Fusion (b) Question-Guided Geometric Memory

Figure 2: **Overview of the proposed framework.** Camera-guided geometry fusion first injects spatial cues into frame tokens. The Fine-Grained Context Bank preserves recent detailed visual evidence, while the Semantic-Geometric Evidence Bank stores compact pooled fused features. Question-conditioned utilities are recomputed from the current question and the active evidence bank, enabling memory-aware evidence reading and replacement.

where $\hat{f}_{h,t} \in \mathbb{R}^{N_v \times d}$. Finally, a camera-conditioned SwiGLU-style channel gate controls how strongly the geometry residual is injected into the visual features:

$$\mathbf{g}_c = \mathrm{SwiGLU}\big(W^g_{c,v} f_{c,t}\big)\,, \tag{3}$$

$$f_{h,t} = f_{v,t} + \mathbf{g}_c \odot W^g_o \hat{f}_{h,t}, \tag{4}$$

where $\mathbf{g}_c \in \mathbb{R}^{1\times d}$ is broadcast over tokens and $f_{h,t} \in \mathbb{R}^{N_v \times d}$. The geometry-aware feature $f_{h,t}$ is shared by both memory branches in Sec. 3.3.

3.3 Q-GeoMem

Q-GeoMem maintains two complementary memory branches over $f_{h,t}$: FGCB for short-range dense context and SGEB for long-range evidence selected by question relevance and novelty.

3.3.1 Fine-Grained Context Bank

The Fine-Grained Context Bank (FGCB) provides short-range dense context by reading recent fused features and camera variation before the current frame is written. At step t, with context window size τ, we maintain a window-based memory:

$$\mathcal{F}_{t-1} = [(f_{h,u}, f_{c,u})]_{u=\max(1,t-\tau)}^{t-1}, \tag{5}$$

where u indexes the recent past frames retained in the buffer.

Memory read. Viewpoint changes play an important role in tasks such as camera displacement or movement direction. To modulate the readout by how the current viewpoint differs from each historical entry, we use u to index each retained past frame and compute a per-entry camera-feature difference that biases the attention logits and gates the memory values in the memory read:

$$\begin{aligned} \Delta f_{c,t,u} &= f_{c,t} - f_{c,u}, \\ b_{t,u} = \mathrm{MLP}^{\mathcal{F}}_b(\Delta f_{c,t,u}), &\quad a_{t,u} = \sigma\big(\mathrm{MLP}^{\mathcal{F}}_a(\Delta f_{c,t,u})\big), \end{aligned} \tag{6}$$

where $b_{t,u}, a_{t,u} \in \mathbb{R}$ are scalars associated with history entry u. The FGCB readout is then computed as camera-modulated attention over the concatenated historical features:

$$R_t^{\mathcal{F}} = \mathrm{Attn}\big(W_q^{\mathcal{F}} f_{h,t},\ W_k^{\mathcal{F}}[f_{h,u}]_{\mathcal{F}_{t-1}},\ (W_v^{\mathcal{F}}[f_{h,u}]_{\mathcal{F}_{t-1}}) \odot [a_{t,u}]_{\mathcal{F}_{t-1}}\ ;\ [b_{t,u}]_{\mathcal{F}_{t-1}}\big)\,, \tag{7}$$

where $R_t^{\mathcal{F}} \in \mathbb{R}^{N_v \times d}$, and the bracketed tensors concatenate all FGCB entries in temporal order. The gate $[a_{t,u}]_{\mathcal{F}_{t-1}}$ rescales the memory values, and the term after the semicolon $[b_{t,u}]_{\mathcal{F}_{t-1}}$ is added to the attention logits as a per-key bias.

Memory write. After reading, FGCB appends $(f_{h,t}, f_{c,t})$ and discards the oldest entry when the buffer exceeds τ. This sliding buffer preserves fine-grained geometry-aware visual features that support local spatial relations and view-dependent details.

3.3.2 Semantic-Geometric Evidence Bank

SGEB stores a fixed-size bank of compact semantic-geometric entries with capacity K. At step t, let $\mathcal{K}_{t-1} \subseteq \{1, \ldots, t-1\}$ denote the indices of past frames currently retained in the bank, with $|\mathcal{K}_{t-1}| \le K$, so that each $u \in \mathcal{K}_{t-1}$ refers to one stored entry. We maintain

$$\mathcal{S}_{t-1} = [(e_u, r_u)]_{u \in \mathcal{K}_{t-1}}, \tag{8}$$

where $e_u \in \mathbb{R}^{M \times d}$ is the pooled semantic-geometric entry extracted from frame u, and $r_u \in \mathbb{R}$ is its calibrated question-relevance score.

Evidence utility. We score each candidate frame by multiplying its question relevance with its novelty relative to an active set, which discourages repeatedly storing redundant evidence. For frame i, we compute its question relevance

$$r_i = \mathrm{Calib}(\mathrm{sim}(\mathcal{Q}(f_{v,i}), \bar{q}))\,, \tag{9}$$

where $\bar{q} \in \mathbb{R}^d$ is the pooled question feature from the text branch of $\mathcal{Q}$, and $\mathrm{Calib}(\cdot)$ is the video-level calibration detailed in Sec. A.2. Relevance is computed per frame, whereas novelty is measured against the active entries used by the current operation. Let $\mathcal{A}$ denote this active index set: $\mathcal{A} = \mathcal{K}_{t-1}$ for memory reading and $\mathcal{A} = \mathcal{C}_t$ for capacity-based replacement. For each $u \in \mathcal{A}$, we define

$$\nu_u(\mathcal{A}) = 1 - \max_{v \in \mathcal{A},\, v \neq u} \mathrm{sim}(e_u, e_v), \tag{10}$$

$$w_u(\mathcal{A}) = r_u\, \nu_u(\mathcal{A}), \tag{11}$$

so that ν_u is large when e_u differs from the other active entries, and w_u favors evidence that is simultaneously question-relevant and non-redundant. When unambiguous, we omit the explicit dependence on $\mathcal{A}$ and write ν_u and w_u.

Memory read. At step t, SGEB reads from the existing bank $\mathcal{S}_{t-1}$ before any update. The current fused feature queries the stored entries, while the evidence utility of each entry is injected as a per-key bias on the attention logits:

$$R_t^{\mathcal{S}} = \mathrm{Attn}\big(W_q^{\mathcal{S}} f_{h,t},\ W_k^{\mathcal{S}}[e_u]_{\mathcal{S}_{t-1}},\ W_v^{\mathcal{S}}[e_u]_{\mathcal{S}_{t-1}}\ ;\ \beta[w_u]_{\mathcal{S}_{t-1}}\big)\,, \tag{12}$$

where $R_t^{\mathcal{S}} \in \mathbb{R}^{N_v \times d}$ and the bracketed tensors concatenate all SGEB entries in temporal order. The term after the semicolon adds the per-key bias $\beta\, w_u$ to the attention logits, where $\beta = \mathrm{softplus}(\theta_h) \ge 0$ is a learnable scale for each w_u. Thus, the evidence utility softly prioritizes high-value entries in the attention logits without directly scaling the memory values.

Memory write. After reading, the candidate entry is presented to the bank. While $|\mathcal{K}_{t-1}| < K$, the candidate is appended. Once the bank is full, we form the candidate set $\mathcal{C}_t = \mathcal{K}_{t-1} \cup \{t\}$, recompute the utility w_u of each entry with novelty measured over $\mathcal{C}_t$, and drop

$$u^\star = \arg\min_{u \in \mathcal{C}_t} w_u. \tag{13}$$

If $u^\star = t$, the candidate is discarded. Otherwise, the lowest-scoring stored entry is evicted and the candidate is inserted. The relevance–novelty product w_u therefore drives both capacity-based replacement and subsequent memory reading (Eq. (12)) without being stored as a static memory attribute.

3.4 Adaptive Fusion

We fuse the memory readouts into the current feature with channel-wise gates conditioned on mean-pooled summaries (denoted by bars, where each visual summary is pooled along the token dimension to $\mathbb{R}^{1\times d}$):

$$\mathbf{g}_t^{\mathcal{F}} = \sigma\big(\mathrm{MLP}_g^{\mathcal{F}}\left([\bar{f}_{h,t}; \bar{R}_t^{\mathcal{F}}]\right)\big), \quad \mathbf{g}_t^{\mathcal{S}} = \sigma\big(\mathrm{MLP}_g^{\mathcal{S}}\left([\bar{f}_{h,t}; \bar{q}; \bar{R}_t^{\mathcal{F}}; \bar{R}_t^{\mathcal{S}}]\right)\big), \\ \tilde{f}_{h,t} = f_{h,t} + \mathbf{g}_t^{\mathcal{F}} \odot R_t^{\mathcal{F}} + \mathbf{g}_t^{\mathcal{S}} \odot R_t^{\mathcal{S}}. \tag{14}$$

where $\bar{q}$ is the projected question summary, $\mathbf{g}_t^{\mathcal{F}}, \mathbf{g}_t^{\mathcal{S}} \in \mathbb{R}^{1\times d}$ are token-broadcast gates, and the fused frame representations are provided to the multimodal LLM.

4 Experiments

4.1 Implementation Details

Training details. Q-GeoMem is built on LLaVA-NeXT-Video-7B (Zhang et al., 2024b) and follows the VLM-3R (Fan et al., 2025) training recipe for video preprocessing, frame sampling, instruction tuning, optimization, and geometry/camera feature extraction. We train on VLM3R-Data (Fan et al., 2025) and VICA-322K (Feng, 2025) for two epochs. For Q-GeoMem-specific hyperparameters, we set the FGCB context window to $\tau = 4$, the SGEB capacity to $K = 8$, and the number of pooled semantic-geometric tokens per frame to $M = 7\times7$. Full training and spatial-tower configurations are provided in Sec. A.1.

Benchmarks and metrics. We conduct a comprehensive evaluation of Q-GeoMem across the main video spatial reasoning setting and several out-of-distribution (OOD) transfer settings. VSI-Bench (Yang et al., 2025a) and VSTI-Bench (Fan et al., 2025) serve as the main video spatial reasoning benchmarks, covering static and temporally evolving spatial questions under camera motion. To assess OOD generalization beyond the main setting, we further evaluate on five complementary benchmarks, including SPBench-MV (Li et al., 2025a) and MMSI-Video (Lin et al., 2025) for video spatial reasoning, SQA3D (Ma et al., 2023) for situated 3D scene question answering, and VideoMME (Fu et al., 2025) and EgoSchema (Mangalam et al., 2023) for general long-form video understanding. We use the multi-view (MV) subset of SPBench, which better matches the multi-view reasoning targeted by Q-GeoMem. VSTI-Bench† denotes the out-of-domain setting in which the model is not trained on the VLM-3R VSTI training set.

Baseline methods. We compare with proprietary VLMs, open-source video VLMs, and spatially specialized models, including VG-LLM (Zheng et al., 2025), VST (Yang et al., 2025b), VLM-3R (Fan et al., 2025), and VLM^2 (Liu et al., 2025), which incorporate geometry, spatial supervision, or memory mechanisms.

4.2 Results

VSI-Bench. As shown in Tab. 1, Q-GeoMem achieves an average score of 71.2, outperforming the strongest spatial-memory baseline VLM^2-7B by 2.4 points and VLM-3R-7B by 10.3 points. Compared with VLM-3R, Q-GeoMem improves all eight task categories, with particularly large gains on appearance order (+34.5), absolute distance (+10.5), room-size estimation (+8.2), and object-size estimation (+7.3). These tasks require aggregating observations from different viewpoints and retaining evidence that may be separated by long temporal intervals, making them sensitive to how memory is written and read. Compared with VLM^2-7B, Q-GeoMem improves six of the eight tasks and remains slightly lower on relative direction and route planning; the gains concentrate on questions requiring complementary evidence across time, whereas route-level planning may require more explicit path reasoning.

VSTI-Bench. For this setting, we additionally fine-tune Q-GeoMem on the VSTI training data and evaluate temporal spatial reasoning under changing camera viewpoints. As shown in Tab. 2, Q-GeoMem obtains an average score of 67.8, improving over VLM^2-7B by 2.5 points and VLM-3R-7B by 9.0 points, with the best results on four of the five task categories. The largest gain over

Table 1: **Evaluation on VSI-Bench (Yang et al., 2025a).** We report average accuracy and eight task-wise scores for proprietary VLMs, open-source video VLMs, and spatially specialized models. Best and second-best results are highlighted within each model category. Q-GeoMem achieves the best average score and shows consistent gains over VLM-3R, indicating that question-guided geometric memory improves the retention of task-relevant spatial evidence.

Methods	Avg.	Numerical Question				Multiple-Choice Question			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models (API)									
GPT-4o (Hurst et al., 2024)	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash (Team et al., 2024)	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro (Team et al., 2024)	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced VLMs									
LLaVA-NeXT-Video-7B (Zhang et al., 2024b)	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B (Zhang et al., 2024b)	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Qwen2.5VL-7B (Bai et al., 2025)	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
LLaVA-OneVision-7B (Li et al., 2024)	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B (Li et al., 2024)	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
InternVL3-78B (Zhu et al., 2025)	48.5	71.2	53.7	44.4	39.5	55.9	39.5	28.9	54.5
Specialized Spatial Reasoning Models									
VG-LLM-8B (Zheng et al., 2025)	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
VST-7B (Yang et al., 2025b)	60.6	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2
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
VLM^2-7B (Liu et al., 2025)	68.8	72.5	59.6	70.8	69.9	69.0	87.8	52.6	68.3
Q-GeoMem (Ours)	71.2	74.8	59.9	76.5	75.3	71.1	87.2	50.0	74.6
Improvement over VLM-3R	+10.3	+4.6	+10.5	+7.3	+8.2	+5.7	+6.7	+4.6	+34.5

Table 2: **Evaluation on VSTI-Bench (Fan et al., 2025).** We compare representative proprietary models, open-source video VLMs, and spatially specialized models on five temporal spatial reasoning tasks. Best and second-best results are highlighted within each model category. Q-GeoMem achieves the best performance, showing the benefit of coupling camera-aware memory readout with question-guided evidence writing.

Methods	Avg.	Numerical Question		Multiple-Choice Question		
		Cam-Obj Abs. Dist.	Cam. Displace.	Cam. Mov. Dir.	Obj-Obj Rel. Pos.	Cam-Obj Rel. Dist.
Proprietary Models (API)						
GPT-4o (Hurst et al., 2024)	38.2	29.5	23.4	37.3	58.1	42.5
Gemini-1.5 Flash (Team et al., 2024)	32.1	28.5	20.9	24.4	52.6	33.9
Open-sourced VLMs						
LLaVA-NeXT-Video-7B (Zhang et al., 2024b)	40.0	28.2	1.8	49.8	64.7	55.6
LLaVA-OneVision-7B (Li et al., 2024)	41.7	29.9	19.3	47.5	62.1	49.8
LongVA-7B (Zhang et al., 2024a)	32.3	13.5	5.1	43.7	57.9	41.2
InternVL2-8B (Chen et al., 2024b)	43.5	32.9	13.5	48.0	68.0	55.0
LongVILA-8B (Chen et al., 2024a)	30.5	20.0	11.6	35.4	52.3	33.4
VILA-1.5-8B (Lin et al., 2024)	37.3	30.1	27.3	42.2	50.4	36.7
VILA-1.5-40B (Lin et al., 2024)	38.2	28.2	15.7	28.8	65.4	53.0
LLaVA-NeXT-Video-72B (Zhang et al., 2024b)	44.0	32.3	10.5	48.1	78.3	50.9
Specialized Spatial Reasoning Models						
VLM-3R-7B (Fan et al., 2025)	58.8	39.4	39.6	60.6	86.5	68.6
VLM^2-7B (Liu et al., 2025)	65.3	43.1	44.1	76.8	87.7	74.9
Q-GeoMem (Ours)	67.8	44.3	44.6	84.0	91.0	74.9

VLM^2-7B appears on camera movement direction (+7.2), a task that directly depends on comparing local viewpoint changes across adjacent observations. This improvement aligns with the FGCB read-modulation ablation in Tab. 5, where Camera-Δ improves FGCB readout by injecting camera variation into local memory retrieval. The VSTI-Bench comparison therefore points to complementary roles for the two memories, with Camera-Δ-aware local memory capturing short-range viewpoint changes and question-guided evidence writing supporting longer-range retention.

Out-of-distribution generalization. To test whether the proposed memory design transfers beyond the main settings, we report OOD results in Tab. 3. All five benchmarks in this evaluation are OOD. SPBench-MV and MMSI-Video test video spatial reasoning, SQA3D tests transfer to situated 3D scene question answering, and VideoMME and EgoSchema test transfer to general video understanding. Adding SGEB consistently improves all five OOD benchmarks, by +3.58 on SPBench-MV, +1.52 on MMSI-Video, +0.95 on SQA3D, +1.19 on VideoMME, and +3.30 on EgoSchema. The largest gains appear on SPBench-MV and EgoSchema, covering both spatially focused transfer and long-form egocentric video QA, where relevant evidence may be sparsely distributed across the sequence.

Table 3: **Out-of-distribution generalization.** We evaluate the effect of SGEB on five OOD benchmarks spanning video spatial reasoning, situated 3D scene QA, and general video understanding. All results use the official metric of each benchmark.

Model	SPBench-MV	MMSI-Video	SQA3D	VideoMME	EgoSchema
Q-GeoMem (w/o SGEB)	69.03	23.97	49.43	59.77	49.81
Q-GeoMem	72.61	25.49	50.38	60.96	53.11

4.3 Ablation Studies

Table 4: **Core component ablation on VSI-Bench.** Starting from the fine-tuned LLaVA-NeXT-Video-7B baseline, we progressively add camera-guided geometry fusion (CGGF), FGCB, and SGEB. Results include average accuracy and task-wise scores for numerical and multiple-choice spatial reasoning.

Model	Avg.	Numerical Question				Multiple-Choice Question			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
LLaVA-NeXT-Video-7B ft. (Baseline)	64.83	71.91	51.47	75.38	70.56	67.32	72.09	42.27	67.64
Baseline + CGGF	67.28	72.07	56.13	74.49	72.43	68.17	81.47	45.36	68.12
Baseline + CGGF + FGCB	67.92	71.70	55.97	75.47	73.75	66.48	84.62	46.91	68.45
Baseline + CGGF + FGCB + SGEB	71.17	74.76	59.89	76.55	75.28	71.13	87.18	50.00	74.60

Core components. We first ablate the core components to verify how each design choice contributes to video spatial reasoning. The VSI-Bench results are summarized in Tab. 4. Adding camera-guided geometry fusion improves the baseline from 64.83 to 67.28, confirming that geometry is more useful when calibrated by camera state before memory construction. Introducing FGCB further raises the average to 67.92, mainly on tasks that benefit from recent fine-grained context, such as relative direction. The full model reaches 71.17 (+6.34 over the baseline), with SGEB yielding especially large gains on appearance order, which depends on retaining complementary evidence over a longer temporal horizon.

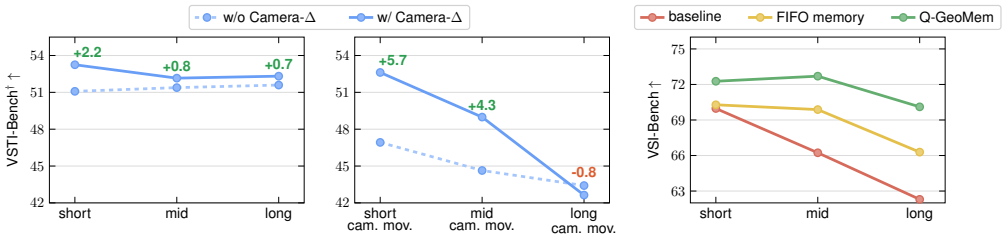
Table 5: **Ablation of FGCB read modulation.** We compare FGCB with and without Camera-Δ while disabling SGEB, and report results by video length on VSI-Bench and VSTI-Bench†.

Model	VSI-Bench				VSTI-Bench†			
	Avg.	Short	Mid	Long	Avg.	Short	Mid	Long
Q-GeoMem (w/o SGEB) w/o Camera-Δ	66.57	68.91	67.66	64.50	51.17	51.08	51.39	51.60
Q-GeoMem (w/o SGEB)	67.92	70.55	69.43	65.33	52.24	53.25	52.16	52.32
Improvement	+1.35	+1.64	+1.77	+0.83	+1.07	+2.17	+0.77	+0.72

Effect of camera modulation. Tab. 5 evaluates whether camera-state variation is a useful conditioning signal for FGCB readout. With SGEB disabled in both variants, the ablated model fixes $b_{t,u} = 0$ and $a_{t,u} = 1$, reducing FGCB reading to standard cross-attention over the recent buffer while keeping the remaining architecture unchanged. Camera-Δ improves VSI-Bench from 66.57 to 67.92 and VSTI-Bench† from 51.17 to 52.24, with positive gains across all length buckets on both benchmarks.

4.4 Analysis

Length-based diagnostics. Fig. 3 examines when local camera modulation and long-range evidence selection are most useful. In panel (a), Camera-Δ improves all length buckets on VSTI-Bench†, most on short videos; on camera movement direction, the gains are clear on short and mid videos but not evident on long videos. This behavior matches the intended division of labor between the two memories. Camera-Δ mainly targets local viewpoint transitions, whereas longer videos require an additional mechanism to select sparse but complementary evidence beyond the recent context window. Panel (b) further compares memory strategies across video length on VSI-Bench. Compared with FIFO memory, Q-GeoMem yields larger gains as video length increases. Long-horizon spatial reasoning benefits from actively selecting question-relevant and non-redundant evidence instead of retaining frames only by chronological order.



(a) VSTI-Bench† across video length and camera motion. **(b)** VSI-Bench across video length and memory strategy.

Figure 3: **Length-based memory diagnostics.** (a) Camera-Δ modulation improves FGCB readout on VSTI-Bench†, especially for short videos and camera movement direction, while its effect is limited on long videos. (b) The proposed SGEB design outperforms FIFO memory on VSI-Bench, with larger gains as the video length increases.

SGEB policy and capacity. Tab. 6(a) compares SGEB replacement and reading policies under a fixed capacity. FIFO reaches 68.7 average accuracy and 66.3 on long videos, while QRel-only and novelty-only scoring provide moderate improvements. The full relevance–novelty utility achieves 71.2 and 70.1, improving over FIFO by +2.5 and +3.8 points and over the best single-term policy by +1.9 points on both metrics. The two terms address complementary failure modes: relevance alone may retain visually similar evidence about the queried content, while novelty alone may preserve diverse but question-irrelevant views.

Table 6: **Policy analysis of SGEB on VSI-Bench.** We vary the SGEB capacity K and compare FIFO replacement, question-relevance-only (QRel-only) scoring, novelty-only scoring, and the full recomputed relevance–novelty utility used by Q-GeoMem.

(a) SGEB Memory Policy

Policy	Update rule	Read bias	VSI Avg. ↑	VSI Long ↑
FIFO	evict oldest	none	68.7	66.3
QRel-only	evict min r	βr	69.3	68.2
Novelty-only	evict min ν	$\beta \nu$	68.9	67.3
Q-GeoMem (Ours)	evict min $r\nu$	$\beta r\nu$	71.2	70.1

(b) SGEB Capacity–Accuracy Trade-off

K	VSI Avg. ↑	VSI Long ↑
4	69.0	67.1
8	71.2	70.1
16	68.4	66.9
24	69.5	68.3
32	67.9	66.2

Tab. 6(b) reveals a non-monotonic capacity–accuracy trade-off. $K = 8$ performs best while keeping only 25% of the 32 sampled frames. A smaller bank ($K = 4$) lacks coverage of complementary evidence, whereas larger banks also underperform, with the full-budget setting $K = 32$ dropping by 3.3 points overall and 3.9 points on long videos. SGEB thus works best as a compact evidence bottleneck that retains diverse observations while penalizing redundant or low-utility entries. The larger gain on the long-video split further reflects the value of compact evidence selection in long-horizon spatial reasoning.

5 Conclusion

We presented Q-GeoMem, a question-guided geometric memory framework for video spatial reasoning that combines camera-conditioned frame calibration with FGCB for recent dense evidence and SGEB for compact long-range evidence selected by question relevance and novelty. Experiments and ablations show consistent gains over strong geometry-aware and memory-based baselines, especially when reasoning requires temporally distributed observations or camera-motion cues. The findings emphasize that effective video spatial reasoning depends not only on stronger visual representations, but also on explicit selection of the spatial evidence retained for each question.

Limitations. Despite these results, Q-GeoMem has several limitations. The framework relies on extracted camera and geometry cues, so inaccurate pose, depth, or reconstruction signals may weaken memory selection. It also does not explicitly model complex multi-step reasoning, such as route-level planning or long chains of spatial transformations. Extending the method to fully online or real-time embodied settings remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools only to aid and polish the writing of this paper, e.g., improving grammar, wording, and readability, and for routine code completion and debugging. We have not used generative AI tools for any task with required disclosure under the ICLR 2027 AI policy. We have reviewed all AI-assisted text and code, and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Sec. 3 describes the full architecture of Q-GeoMem, and Sec. 4 specifies the base model, training data, hyperparameters, benchmarks, and evaluation protocols. Sec. A provides the complete training configuration and the exact definitions of video-level calibration and the active evidence utility. All training data and benchmarks are publicly available. Our anonymized code, including training and evaluation scripts, is available at https://github.com/qgeomem/Q-GeoMem, and the trained checkpoint at https://huggingface.co/qgeomem/q-geomem-llava-qwen2-vsi. Our anonymous project page at https://qgeomem.github.io further provides dynamic demos of Q-GeoMem.

REFERENCES

Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. HierarQ: Task-aware hierarchical Q-former for enhanced video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8545–8556, 2025.

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:2502.13923*, 2025.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082, 2025.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. *arXiv:2408.10188*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024b.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. ReWind: Understanding long videos with instructed learnable memory. *arXiv:2411.15556*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. *arXiv:2505.20279*, 2025.

Qi Feng. Visuospatial cognitive assistant. *arXiv:2505.12312*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv:2405.21075*, 2025.

Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, and Bin Zhao. Learning 2D Invariant Affordance Knowledge for 3D Affordance Grounding. *arXiv:2408.13024*, 2025.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712, 2025.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-online. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338, 2025.

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

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv:2506.03135*, 2026.

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

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: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:2510.08531*, 2025a.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs Ready for Precise Spatial-Temporal World Understanding? *arXiv:2503.23765*, 2025b.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. *arXiv:2312.07533*, 2024.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, Wenbo Hu, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Video-Bench: A Holistic Benchmark for Video-Based Spatial Intelligence. *arXiv:2512.10863*, 2025.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-Language Memory for Spatial Reasoning. *arXiv:2511.20644*, 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. *arXiv:2210.07474*, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *arXiv:2308.09126*, 2023.

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:2504.01805*, 2025.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24045–24055, 2025.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv:2604.02317*, 2026.

Haoming Song, Delin Qu, Yuanqi Yao, Qizhi Chen, Qi Lv, Yiwen Tang, Modi Shi, Guanghui Ren, Maoqing Yao, Bin Zhao, et al. Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive Keyframe Sampling for Long Video Understanding. *arXiv:2502.21271*, 2025.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *arXiv:2501.12387*, 2025a.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. VideoLLaMB: Long Streaming Video Understanding with Recurrent Memory Bridges. *arXiv:2409.01071*, 2025b.

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. *arXiv:2501.13468*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 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:2511.05491*, 2025b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. *arXiv:2505.23764*, 2025c.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21059–21069, 2025.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. LLaVA Blog, https://llava-vl.github.io/blog/2024-04-30-llava-next-video, 2024b.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models. *arXiv:2511.23075*, 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors. *arXiv:2505.24625*, 2025.

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.

A Implementation Details

A.1 Training Configuration

Q-GeoMem is trained with LoRA fine-tuning using rank 128 and alpha 256. We use per-device batch size 1, gradient accumulation steps 8, learning rate 2×10^{-5}, weight decay 0, warmup ratio 0.03, and a cosine learning-rate schedule. Training uses BF16 and TF32 precision, DeepSpeed ZeRO stage 2, and gradient checkpointing. Following VLM-3R (Fan et al., 2025), we use CUT3R (Wang et al., 2025a) as the spatial tower, select all spatial features with feature dimension 768, keep the spatial tower frozen, and tune the fusion block and multimodal MLP adapter. Training takes approximately 38 hours on 16 H200 GPUs.

A.2 Video-Level Calibration

The video-level calibration $\mathrm{Calib}(\cdot)$ used in the question-relevance score of Eq. (9) standardizes the raw question–frame similarity within each video, so that relevance is comparable across videos whose overall similarity scales differ. For frame i, let $\rho_i = \mathrm{sim}(\mathcal{Q}(f_{v,i}), \bar{q})$ denote the raw similarity between the Q-Former frame embedding and the pooled question feature. We calibrate it as

$$r_i = \sigma\left(\frac{\rho_i - \mu_\rho}{\sigma_\rho + \epsilon}\right),$$

where μ_ρ and σ_ρ are the mean and standard deviation of $\{\rho_i\}$ computed over the sampled frames of the video, $\epsilon > 0$ is a small constant for numerical stability, and $\sigma(\cdot)$ is the logistic sigmoid that maps the standardized score to $(0, 1)$. This per-video standardization removes video-specific similarity offsets and scales, which stabilizes the relevance–novelty utility used for SGEB reading and replacement.

A.3 Active Evidence Utility

As described in Sec. 3.3, SGEB recomputes the relevance–novelty utility over the active entries used by each operation, rather than storing it as a fixed scalar. For read-time retrieval, the active set is the current bank before the candidate frame is written; for replacement, it is the full candidate set after adding the current frame. This distinction ensures that retrieval reflects the evidence currently available, while eviction accounts for the redundancy introduced by the new candidate. When the active set contains only one entry, we set its novelty to 1, so the first retained entry is scored solely by question relevance. The resulting utility is then used consistently as the read-time attention bias in Eq. (12) and as the replacement score under the fixed memory budget.

B Additional Results

B.1 Evidence Utility Usage

Tab. 7 isolates where the full relevance–novelty utility $w = r\nu$ should be used in SGEB. The retrieval-off variants remove the utility bias in read attention, while the replacement-off variants retain chronological FIFO eviction. This 2-by-2 design reports VSI-Bench average accuracy and tests whether the same active utility should jointly control evidence retrieval and memory replacement.

Table 7: **Disentangling evidence utility usage.** Starting from the FIFO memory baseline, we fix the full relevance–novelty utility $w = r\nu$ and enable it either for read-time attention bias, for capacity-based replacement, or for both. We report average accuracy on VSI-Bench, isolating whether the gain comes from better retrieval, better memory composition, or their combination.

Variants	Active w for replacement	Read bias βw	VSI Avg.
FIFO	✗	✗	68.7
Read bias only	✗	✓	69.2
Write replacement only	✓	✗	70.2
Q-GeoMem (Ours)	✓	✓	71.2

B.2 Additional Qualitative Results

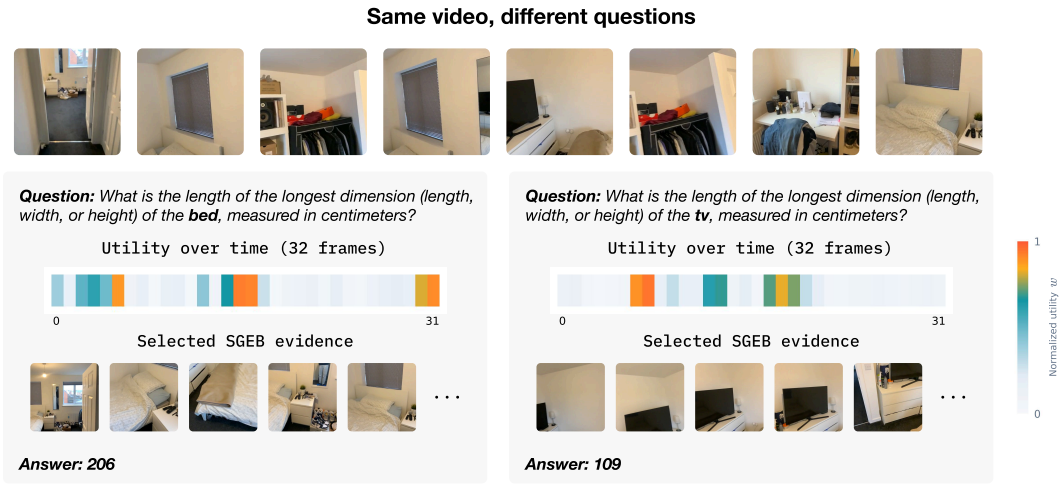


Figure 4: **Question-conditioned SGEB evidence selection.** The heatmaps show normalized active SGEB utility $w = r\nu$ over 32 frames for two questions from the same video, and the selected entries indicate the evidence retained by the bank.

To illustrate how SGEB adapts evidence selection to the question, we visualize qualitative examples in Fig. 4. Although the visual stream is identical, changing the queried object from the bed to the TV shifts both the high-utility frames and the retained entries. For the bed question, Q-GeoMem preserves complementary views of the bed and nearby context, whereas for the TV question, the retained evidence shifts toward frames where the TV is visible. The retained bank is therefore not a fixed chronological cache, but a question-conditioned evidence store that prioritizes relevant and non-redundant observations for the current spatial query.

We provide additional qualitative examples in Fig. 5 to further visualize how SGEB selects question-conditioned evidence from the same video.



Figure 5: **Additional question-conditioned SGEB evidence selection from the same video.**