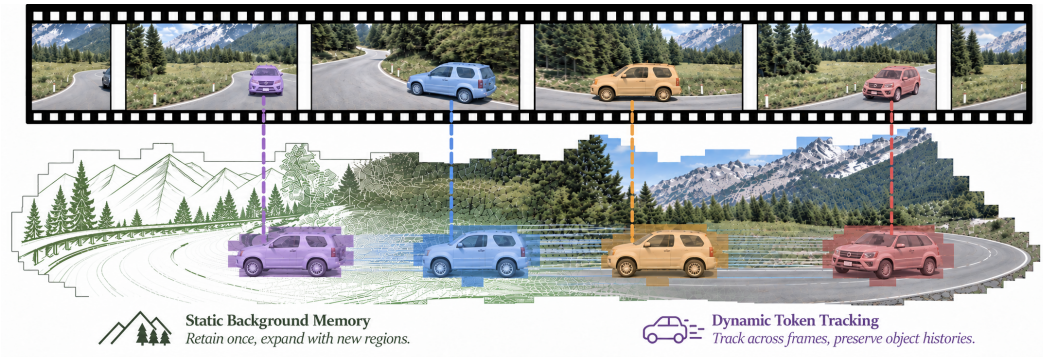
EGOWORLDMEM: 4D WORLD MEMORY FOR LONG-HORIZON EGOCENTRIC VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric videos contain substantial redundancy as the same physical environment is repeatedly observed from changing viewpoints, while brief object interactions often carry critical evidence. Existing memory mechanisms typically compress visual history based on temporal proximity or feature similarity, making it difficult to distinguish camera motion from actual changes in the world. We introduce **EgoWorldMem**, a 4D world-memory framework that compacts redundant observations of the physical world rather than merely historical tokens. EgoWorldMem reprojects an evolving geometric map into each incoming view to compensate for ego motion, retains sparse world-grounded visual evidence, and organizes dynamic observations into object-centric event segments through persistent tracking. The resulting memory is serialized into a compact visual-token sequence for VLM inference. Across five video-understanding benchmarks, EgoWorldMem improves the average accuracy of Qwen3-VL and LLaVA-NeXT-Video by 10.0 and 23.6 percentage points while removing 73.96% and 78.65% of visual tokens, respectively. An online variant further reduces time-to-first-token by 57.39% on EgoSchema, demonstrating the effectiveness of world-grounded memory for efficient long-horizon egocentric video understanding.



1 INTRODUCTION

Egocentric video understanding is increasingly moving toward always-on assistants that reason over continuous first-person streams. These streams typically unfold within a coherent physical environment, where the wearer repeatedly observes the same places and objects from changing viewpoints. This spatial continuity creates substantial cross-view redundancy while making it difficult to distinguish changes caused by camera motion from changes in the world itself. Such assistants must remember places and objects across time and answer queries that depend on what the wearer has seen, manipulated, or changed, even when the relevant evidence is spatially small and temporally fleeting. A useful system therefore needs memory that is long-horizon enough to preserve past evidence, selective enough to remain efficient, and structured enough to distinguish persistent scene context from transient object motion.

Large video-language models (VLMs) offer a promising foundation for reasoning over long-horizon egocentric streams, but processing an ever-growing video stream is prohibitively expensive. Dense visual tokens increase both computation and memory consumption as video duration grows. Existing systems address this problem through frame sampling, recurrent or hierarchical memory, and

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.

visual-token selection (Chen et al., 2024; Wang et al., 2024b; Song et al., 2024; 2025; Yao et al., 2025; Yang et al., 2025b; Xie et al., 2026). These approaches effectively reduce redundancy in a video sequence, yet their memories are commonly indexed by time or represented as latent summaries. Consequently, they have no explicit mechanism to tell whether a visual difference is caused by camera motion, a newly observed part of a static scene, or an object that has actually moved. This ambiguity is costly for egocentric tasks in which the important evidence is often a brief hand-object interaction embedded in a long, visually repetitive stream.

We argue that the goal is not merely to compress historical tokens, but to compact redundant observations of the world. Recent feedforward geometry models can recover point maps, camera poses, and correspondences from unconstrained visual input (Wang et al., 2024a; 2025a;b), and are increasingly capable of modeling dynamic scenes (Zhang et al., 2026; Fang et al., 2026; Hu et al., 2026). This makes it practical to maintain a world-grounded memory while a video is observed. Such a memory should preserve stable scene evidence once, retain temporally coherent evidence for dynamic objects, and expose both forms of evidence in a representation that a VLM can reason over.

To this end, we present **EgoWorldMem**, a 4D world-memory framework for efficient egocentric video understanding. Rather than compressing visual tokens purely along the temporal dimension, EgoWorldMem organizes observations around the evolving physical world. It maintains an accumulated geometric map and, for each incoming frame, reprojects this map into the current view to establish spatial correspondence across viewpoints. Patch-wise comparison between the reprojected map and the current observation compensates for ego motion, allowing EgoWorldMem to distinguish redundant content from static background and dynamic object candidates. It sparsely retains visual tokens from static background candidates, while using SAM2 to segment and track objects in dynamic object candidates. Persistent object tracks link observations of the same object over time, enabling the geometric map to replace outdated object geometry with current observations and the visual memory to organize dynamic evidence into temporally coherent event segments. The resulting sparse background tokens and object-centric event segments together constitute a compact 4D world memory, which is serialized into a visual-token sequence for direct VLM inference. EgoWorldMem uniquely combines three properties. *Geometric*: it uses an evolving 3D map to compare observations across viewpoints and ground visual tokens in a shared coordinate frame. *Semantic*: it organizes visual evidence by object identity, connecting observations of the same object across frames. *Dynamic*: it preserves the temporal evolution of object interactions while suppressing redundant static observations. Together, these properties turn egocentric video into compact, VLM-ready memory for long-horizon reasoning.

We summarize our contributions as follows: **(1)** We propose EgoWorldMem, to the best of our knowledge the first visual memory framework for egocentric video understanding that compacts redundant static observations and preserves dynamic object histories in a unified world representation. **(2)** EgoWorldMem organizes retained evidence into localized, object-centric event segments, highlighting interactions for VLM reasoning. **(3)** We integrate EgoWorldMem into two backbones, Qwen3-VL and LLaVA-NeXT-Video. Across five benchmarks, our method improves average accuracy by 10.0 and 23.6 percentage points using only 26.04% and 21.35% of the original visual tokens, respectively. We further implement an online variant that constructs memory at 1.3 fps and reduces model time-to-first-token (TTFT) by 57.39%.

2 Related Work

2.1 Egocentric Video Understanding

Egocentric video understanding requires reasoning from a first-person stream where viewpoint changes, hand-object interactions, and long-horizon procedures are tightly coupled. A growing set of benchmarks, including EPIC-KITCHENS (Damen et al., 2020), Ego4D (Grauman et al., 2022), EgoSchema (Mangalam et al., 2023), and VidEgoThink (Cheng et al., 2024), reveals that conventional VLMs still struggle with egocentric reasoning. Subsequent methods improve first-person understanding from different angles: ProAssist (Zhang et al., 2025b) introduces proactive dialogue capabilities, Ego-R1 (Tian et al., 2025) applies tool-augmented reasoning to ultra-long videos. Despite these advances, most methods still operate on sampled frames or compressed latent tokens, with little awareness of the 3D environment surrounding the wearer; EgoWorldMem instead grounds its memory in a 4D world, enabling a more effective allocation of its memory budget.

2.2 Memory Mechanisms for Video-Language Models

Video-language models retain historical context through compressed memories. MovieChat (Song et al., 2024) and VideoLLaMB (Wang et al., 2024b) consolidate observations through long-term buffers and recurrent memory bridges, respectively. StreamMem (Yang et al., 2025b) maintains a bounded KV cache, while FluxMem (Xie et al., 2026) and StreamForest (Zeng et al., 2026) organize history into memory hierarchies and event trees. These methods primarily exploit temporal structure, feature similarity, or attention. Geometry-aware selection additionally accounts for viewpoint changes and spatial coverage: EgoPrune (Li et al., 2025) aligns adjacent frames through homographies before filtering redundant tokens, while Geo3DPruner (Li et al., 2026) selects representative features within and across 3D voxels. These approaches motivate using spatial correspondence to allocate visual evidence under limited token budgets. Explicit spatial memories further connect accumulated observations to VLM reasoning. 3D-Mem (Yang et al., 2025c) stores informative multi-view snapshots, GSMem (Lu et al., 2026) uses 3D Gaussian Splatting to support recollection through novel-view rendering, and Reasmory (He et al., 2026) enables structured queries over reconstructed 3D memory. EgoWorldMem focuses on preserving evidence of an evolving scene: map reprojection guides sparse static retention, persistent object identities organize dynamic observations into event histories, and memory readout serializes both into compact visual tokens for video understanding.

2.3 Feedforward 3D Reconstruction

Feedforward 3D reconstruction has become a scalable alternative to optimization-heavy geometry pipelines. DUSt3R (Wang et al., 2024a) predicts dense point maps for unconstrained images, and VGGT (Wang et al., 2025a) and VGGT-Ω (Wang et al., 2026) unify camera estimation, depth, and point tracking in visual geometry transformers. The paradigm is also moving beyond static scenes: CUT3R (Wang et al., 2025b) handles streaming video with a persistent recurrent state, and recent work recovers dynamic 4D structure from monocular video (Zhang et al., 2026; Fang et al., 2026; Hu et al., 2026). These advances make it practical to lift egocentric streams into temporally aligned 3D structure, directly supporting EgoWorldMem by caching static scene geometry once and retaining dynamic content as motion-aware tokens.

3 Method

The core idea of EgoWorldMem is to compress redundant observations of the world rather than the temporal sequence of visual tokens. This is especially important for egocentric video, which unfolds in a single physical space that the wearer continually revisits. We first define the 4D world-memory representation in Section 3.1. Its construction and readout follow the four stages in Fig. 1: geometric reprojection identifies redundant regions, static background candidates, and dynamic object candidates (Section 3.2); semantic object tracking establishes persistent identities across frames (Section 3.3); the geometric map is updated and world-grounded visual tokens form sparse static evidence and dynamic event segments (Section 3.4); and memory readout serializes these tokens for VLM inference. Finally, Section 3.5 presents an online variant that incrementally builds memory as video arrives, reducing TTFT so that the assistant can respond promptly as the wearer acts.

3.1 4D World Memory Representation

EgoWorldMem accumulates selected visual evidence over time in a 4D world memory, with each retained token grounded in a shared world coordinate frame:

$$\mathcal{M}_t = \{(f_i, \tau_i, \mathbf{x}_i, o_i)\}_{i \in \mathcal{I}_t}, \tag{1}$$

where $\mathcal{I}_t$ indexes the retained token evidence up to time t, f_i is a visual token feature, $\tau_i \leq t$ is its observation time, $\mathbf{x}_i = (x_i, y_i, z_i)$ is its 3D space location at that time, and o_i records its object association, with -1 denoting unassigned background evidence. The time and position together form a 4D anchor (τ_i, x_i, y_i, z_i). Static evidence is stored sparsely, while dynamic evidence is organized into object-centric event segments. Both share the same grounding.

An auxiliary geometric map maintains an up-to-date representation of the scene to support redundancy estimation and token grounding. We leverage a feedforward 3D model $\mathcal{G}$ (Wang et al., 2025a;

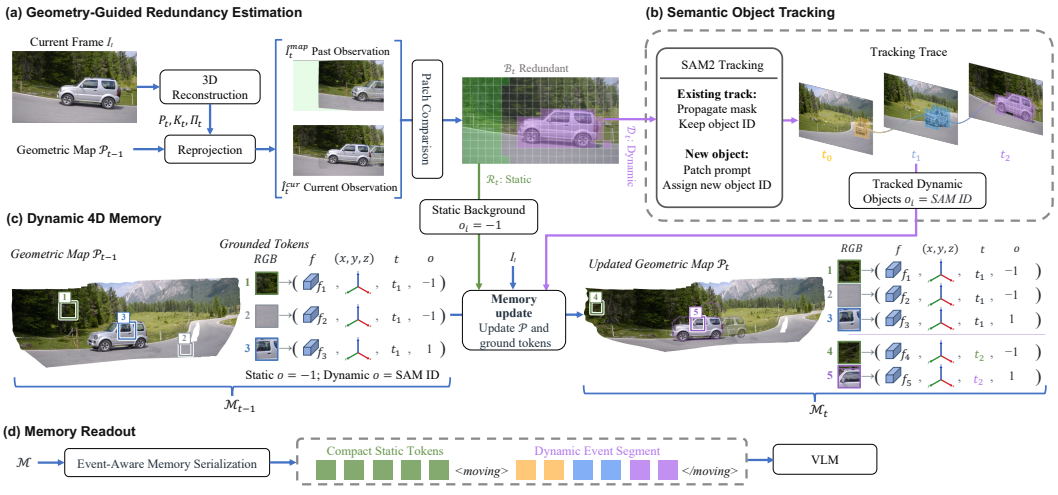


Figure 1: **Overview of EgoWorldMem. (a) Geometry-Guided Redundancy Estimation.** Geometric reprojection compensates for camera motion to identify redundant regions, static background candidates, and dynamic object candidates. **(b) Semantic Object Tracking.** SAM2 propagates object masks and maintains persistent object identities across frames. **(c) Dynamic 4D Memory.** The geometric map is updated, and retained visual tokens f are bound to 3D locations (x, y, z), timestamps t, and object identities o, preserving sparse static evidence and dynamic object histories. **(d) Memory Readout.** Event-aware serialization combines compact static tokens and object-centric dynamic event segments into a visual-token sequence for VLM inference.

2026; 2025b) to process each frame I_t of the egocentric video stream, predicting a dense world-coordinate point map $P_t \in \mathbb{R}^{H\times W\times 3}$, together with camera intrinsics K_t, camera extrinsics Π_t, and per-pixel confidence C_t. Here H and W denote the frame height and width. At time t, the geometric map is represented as

$$\mathcal{P}_t = \{(\mathbf{x}_j, c_j, s_j, \tau_j, o_j)\}_{j=1}^{N_t}, \tag{2}$$

where N_t is the number of geometric map entries, each storing a 3D position $\mathbf{x}_j$, color c_j, reconstruction confidence s_j, observation time $\tau_j \le t$, and object identity o_j. Only observations with valid reconstruction confidence are incorporated into the map. Here $o_j = -1$ marks background, while tracked objects receive distinct identities through the tracking procedure in Section 3.3.

3.2 Geometry-Guided Redundancy Estimation

EgoWorldMem compares the current observation with the mapped scene to guide evidence retention and identify new or changing evidence. In egocentric video, camera motion shifts the same static surface across image patches, making temporal proximity or feature similarity alone insufficient to distinguish repeated observations from scene changes. We reproject the geometric map into the current view and assess changes against the geometrically aligned observations.

Concretely, we reproject $\mathcal{P}_{t-1}$ into frame t using (K_t, Π_t) and resolve occlusions with a z-buffer, obtaining a sparse rendering $\hat{I}_t^{\text{map}}$ of the mapped scene, together with a binary coverage mask O_t that marks the pixels it explains. The reconstructed points of the current frame are projected into the same view to obtain $\hat{I}_t^{\text{cur}}$, and holes in both renderings are inpainted for direct comparison.

We partition both renderings into patches aligned with the VLM's visual token grid and compute appearance discrepancy and map coverage for each patch u:

$$\Delta_t(u) = \sqrt{\frac{1}{|u|}\sum_{i\in u}\|\hat{I}_t^{\text{cur}}(i) - \hat{I}_t^{\text{map}}(i)\|^2}, \qquad \Omega_t(u) = \frac{1}{|u|}\sum_{i\in u} O_t(i). \tag{3}$$

Here, $\Delta_t(u)$ measures the appearance discrepancy between the two renderings within patch u, while $\Omega_t(u)$ quantifies the fraction of the patch covered by projected map points. The two scores jointly

induce a three-way partition:

$$u \in \begin{cases} \mathcal{R}_t & \text{if } \Omega_t(u) < \tau_\Omega \quad \text{(static background candidate)}, \\ \mathcal{B}_t & \text{if } \Omega_t(u) \geq \tau_\Omega \text{ and } \Delta_t(u) < \tau_\Delta \quad \text{(redundant region)}, \\ \mathcal{D}_t & \text{if } \Omega_t(u) \geq \tau_\Omega \text{ and } \Delta_t(u) \geq \tau_\Delta \quad \text{(dynamic object candidate)}. \end{cases} \tag{4}$$

where τ_Δ and τ_Ω are hyperparameters, with their values detailed in Appendix A.1. We denote the corresponding patch-level region masks collectively by $\mathcal{J}_t$, as shown in Fig. 1. Dynamic object candidates $\mathcal{D}_t$ provide candidates for object tracking (Section 3.3), while static background candidates $\mathcal{R}_t$ guide static evidence retention (Section 3.4). We analyze sensitivity to τ_Δ in our ablation studies.

3.3 Semantic Object Tracking

Geometric comparisons identify dynamic object candidates, while object tracking associates these regions across frames. We use SAM2 (Ravi et al., 2025) to propagate existing object masks while preserving their identities. For dynamic object candidates in $\mathcal{D}_t$ not sufficiently covered by existing tracks, we use their patch centers as positive point prompts to generate new object masks and assign new identities. The resulting dynamic object tracking traces link observations of the same object across changes in viewpoint and position, supporting object-wise geometric map updates and the organization of dynamic observations into event histories.

3.4 Dynamic 4D Memory

Evidence for egocentric reasoning often lies in brief interactions, such as picking up or moving an object. To preserve their temporal coherence, EgoWorldMem uses the object tracks to organize dynamic evidence into localized event segments and combines them with sparse static tokens. Both share the same world grounding and together form the memory consumed by the VLM.

Geometric map update. For each tracked object, we aggregate appearance discrepancy and map coverage over its associated patches to assess whether its stored geometry is outdated. When an object's geometry is deemed outdated, we remove its stored points from $\mathcal{P}_{t-1}$. We then incorporate valid observations from the current frame and apply object-wise voxel merging when the map exceeds its size threshold (Appendix A.3), yielding $\mathcal{P}_t$. The detailed map update procedure and its thresholds are provided in Appendix A.2. This update removes outdated geometry from the auxiliary map without discarding historical visual evidence from $\mathcal{M}_t$.

World Memory Grounding. For each candidate visual token f_i produced by the vision encoder, we assign a 3D location $\mathbf{x}_i$ by averaging valid reconstructed points in P_{τ_i} over its source patch. Together with its observation time τ_i, this location forms the token's 4D anchor. A token is associated with an object track when the overlap ratio between its patch and the object's mask reaches a predefined threshold, yielding the object association o_i; otherwise, $o_i = -1$. This binding produces candidate entries $(f_i, \tau_i, \mathbf{x}_i, o_i)$, from which the following retention strategies construct $\mathcal{M}_t$.

Static Evidence Retention. For tokens outside dynamic event segments, we retain evidence only where the geometric map cannot yet explain the view:

$$\mathcal{K}_t = \{u \in \mathcal{R}_t \mid \Delta_t(u) > \tau_\Delta \ \vee \ \Omega_t(u) < \epsilon\}, \tag{5}$$

where $\mathcal{K}_t$ contains static background candidates retained as static evidence: patches with large appearance discrepancies or near-zero map coverage. We then remove small isolated components from the selected patches using the setting in Appendix A.1. This selection retains static evidence from previously unexplained regions while excluding redundant observations of revisited scene content.

Dynamic Event Segmentation. We assess each tracked object's dynamic state by aggregating appearance discrepancy and map coverage over its associated patches. An object is classified as dynamic in a frame when both scores reach τ_Δ and τ_Ω, respectively. We group the dynamic portions of each track into temporal spans and retain spans containing at least $n_{\min}$ associated tokens. Each event segment consists of an object identity, a time interval $[t_s, t_e]$, and the object's tokens within that interval. The overlap threshold and $n_{\min}$ are specified in Appendix A.1. These segments preserve temporally coherent, object-specific evidence, while the static retention rule applies to tokens outside the retained segments. Together, the selected static tokens and dynamic event tokens form $\mathcal{M}_t$.

Memory Readout. EgoWorldMem serializes the two complementary forms of evidence in $\mathcal{M}_t$: sparse static scene observations and object-centric dynamic event segments. Static tokens retain their associations with the original video frames, while each dynamic event is serialized as a contiguous token block at the position of its end frame, enclosed by the markers <moving> and </moving>. Within each retained frame or event segment, visual tokens preserve their original spatial order. For Qwen3-VL, we additionally insert the event's start time into the event block and reconstruct the M-RoPE position IDs from the original spatial positions of the source patches. The readout replaces the original video tokens during prefill without modifying the model architecture.

3.5 Online Memory Construction

We implement a streaming variant of EgoWorldMem by instantiating $\mathcal{G}$ with streaming 3D reconstruction models, such as CUT3R (Wang et al., 2025b), together with SAM2 for online object tracking. The redundancy criterion, evidence retention, and geometric map update use only current and past observations, allowing the 4D world memory to be constructed incrementally while the video is being recorded. This shifts memory construction from query time to video ingestion time. When a query arrives, the VLM directly consumes the compact memory readout without re-encoding the full video history, substantially reducing the time to first token (TTFT). Table 4 reports construction cost and query-time efficiency relative to full-token inference.

4 Experiments

4.1 Experimental Setups

Implementation details. In the main experiments, we construct the geometric map using VGGT-Ω (Wang et al., 2026) and perform object tracking with SAM2 (Ravi et al., 2025). For the online variant, we use CUT3R (Wang et al., 2025b) for streaming 3D reconstruction. During prefill, the compact memory readout is used in place of the original video-token sequence without modifying the VLM architecture. We evaluate EgoWorldMem with two different backbones, Qwen3-VL (Bai et al., 2025) and LLaVA-NeXT-Video (Zhang et al., 2024), to demonstrate its effectiveness across VLM architectures. Detailed hyperparameter settings are provided in Appendix A.1.

Training datasets. We fine-tune both VLM backbones on NExT-QA and QaEgo4Dv2 training splits sampled at 1 fps, together with 1K QA examples constructed from EgoIT videos sampled at 4 fps following VLM4D's question format (Appendix C.3). This adaptation helps the models accommodate the reorganized token sequences produced by our memory readout. The combined training set contains 120K samples spanning 339 hours of egocentric video. We also evaluate a training-free variant that preserves the original token-sequence structure during memory readout. As shown in Table 5, it consistently outperforms random token retention under the same token budget, demonstrating that our geometry-grounded selection retains more informative visual evidence even without adaptation.

Baselines. We compare EgoWorldMem against three groups of baselines. The proprietary API-based models include GPT-4o (Hurst et al., 2024), Gemini-1.5 Pro (Team et al., 2024), and the Qwen3.8 series (Qwen Team, 2026c;b). The open-source models include LLaVA-NeXT-Video (Zhang et al., 2024), Qwen3-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026a), InternVideo3 (Yan et al., 2026), and InternVL3.5 (Wang et al., 2025c). The memory-augmented VLMs include LiveVLM (Ning et al., 2025), Flash-VStream (Zhang et al., 2025a), StreamForest (Zeng et al., 2026), and FluxMem (Xie et al., 2026). We use 32 uniformly sampled frames for the API-based models, except GPT-4o and Gemini-1.5 Pro, whose results are taken from prior work under their respective evaluation settings. The remaining open-source models are evaluated at 1 fps. For the memory-augmented VLMs, we adjust their configurations to match the compression rate of EgoWorldMem with Qwen3-VL, ensuring a fair comparison under comparable token budgets. Detailed baseline configurations are provided in Appendices B.2 and B.3.

Benchmarks. We evaluate EgoWorldMem on five video-understanding benchmarks: VLM4D (Zhou et al., 2025), EgoSchema (Mangalam et al., 2023), QaEgo4Dv2 (Bärmann &

Table 1: **Main results of EgoWorldMem and baselines on five benchmarks.** Overall denotes mean accuracy or pooled token reduction (green). Appendix C details the metrics and benchmarks. Bold and underlined values indicate the best and second-best accuracy results, respectively.

Model	VLM4D	EgoSchema	QA-Ego4D	NextQA	NextGQA	Overall
API-based Models						
GPT-4o† (Hurst et al., 2024)	57.5	67.8	26.3	71.2	–	–
Gemini-1.5 Pro† (Team et al., 2024)	–	72.2	50.8	70.4	–	–
Qwen3.8-Max (Qwen Team, 2026c)	65.4	69.5	49.5	77.7	72.9	67.0
Qwen3.8-Flash (Qwen Team, 2026b)	61.9	64.3	52.3	75.7	70.0	64.8
Open-Source Models						
LLaVA-NeXT-Video-7B (Zhang et al., 2024)	35.2	41.5	45.5	65.2	59.5	49.4
LLaVA-NeXT-Video-34B (Zhang et al., 2024)	37.5	49.4	46.7	70.5	66.2	54.1
Qwen3-VL-8B-Instruct (Bai et al., 2025)	56.0	71.8	63.9	81.9	79.1	70.5
Qwen3-VL-32B-Instruct (Bai et al., 2025)	65.0	<u>77.0</u>	64.9	84.4	82.3	74.7
Qwen3.5-9B (Qwen Team, 2026a)	59.0	66.2	<u>68.5</u>	58.6	79.5	66.4
Qwen3.5-27B (Qwen Team, 2026a)	<u>69.1</u>	<u>77.0</u>	**70.9**	83.6	82.8	<u>76.7</u>
InternVideo3-8B (Yan et al., 2026)	40.3	63.9	47.3	64.0	63.2	55.7
InternVL3.5-8B (128f) (Wang et al., 2025c)	52.5	61.1	57.5	81.3	79.8	66.4
InternVL3.5-38B (128f) (Wang et al., 2025c)	56.3	68.5	46.7	84.4	66.2	64.4
Memory-Augmented VLMs						
LiveVLM (Ning et al., 2025)	43.8	57.7	51.3	77.2	74.3	60.9
Flash-VStream (Zhang et al., 2025a)	40.0	60.8	50.9	76.5	72.2	60.1
StreamForest (Zeng et al., 2026)	49.3	57.2	56.7	81.1	77.0	64.3
FluxMem (Xie et al., 2026)	46.6	59.2	56.1	71.4	75.3	61.7
EgoWorldMem (Ours)						
EgoWorldMem (LLaVA-NeXT-Video)	66.2	71.7	57.7	<u>85.5</u>	<u>84.0</u>	73.0 (+23.6)
Compression (%) ↑	71.36	78.05	71.11	81.52	80.76	78.65
EgoWorldMem (Qwen3-VL)	**74.9**	**83.6**	<u>68.5</u>	**88.6**	**87.1**	**80.5** (+10.0)
Compression (%) ↑	63.46	73.38	67.04	77.01	76.26	73.96

† Results for GPT-4o and Gemini-1.5 Pro are reported in prior work; we detail the baseline sources and input settings in Appendix B.2.

Waibel, 2022; Di & Xie, 2024; Patel et al., 2025), NExT-QA (Xiao et al., 2021), and NExT-GQA (Xiao et al., 2024). Within each video, events unfold in a shared physical space, consistent with the spatial continuity modeled by EgoWorldMem. All five benchmarks adopt a multiple-choice question-answering format, and results are reported as accuracy.

4.2 Main Results

Table 1 reports the comparison across five benchmarks, with a visual summary in Fig. 2. The results show that EgoWorldMem improves performance on both VLM backbones while consolidating visual evidence into a compact world memory. With Qwen3-VL, EgoWorldMem achieves the highest average accuracy of 80.5 at a visual-token compression rate of 73.96%, outperforming its Qwen3-VL-8B backbone by 10.0 percentage points and the strongest competing model, Qwen3.5-27B, by 3.8 points. The gain is larger for LLaVA-NeXT-Video-7B, whose baseline accuracy is lower: EgoWorldMem increases its average accuracy from 49.4 to 73.0, an improvement of 23.6 points, at a visual-token compression rate of 78.65%. We examine the cost of memory construction and its query-time benefits in Table 4.

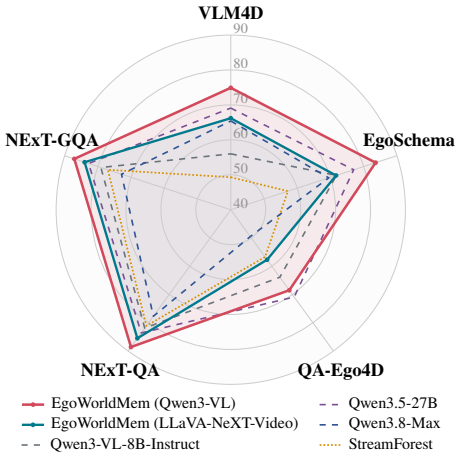


Figure 2: Performance comparison across five video-understanding benchmarks.

EgoWorldMem also outperforms all evaluated memory-augmented VLMs. Compared with StreamForest (Zeng et al., 2026), an advanced training-based memory VLM, EgoWorldMem improves average accuracy by 8.7 percentage points with LLaVA-NeXT-Video and 16.2 points with Qwen3-VL, outperforming it on all five benchmarks with both backbones. These results support the effectiveness of EgoWorldMem in preserving task-relevant visual evidence while compacting redundant observa-

Table 2: **Component ablations on Qwen3-VL-8B.** All variants follow the same fine-tuning protocol with aligned visual-token budgets. Avg. denotes mean accuracy across the five benchmarks.

Memory configuration	Accuracy (%)					
	VLM4D	EgoSchema	QA-Ego4D	NextQA	NextGQA	Avg.
Random token retention	65.6	80.8	58.7	80.7	79.9	73.1
EgoWorldMem w/o event segments	69.6	83.0	67.5	88.0	86.5	78.9
EgoWorldMem	74.9	83.6	68.5	88.6	87.1	80.5

tions. Tracked dynamic-object token sequences further help the model capture brief but important motion cues, supporting egocentric video understanding.

4.3 Ablation Studies

We examine the contributions of geometry-guided evidence selection and object-centric event organization through controlled comparisons in Table 2. We implement two variants: random token retention and EgoWorldMem without event segments. The former randomly selects visual tokens from the full video sequence to match EgoWorldMem's per-video token budget, while the latter uses geometry-guided retention. Both variants keep retained tokens in the frame-wise sequence without extracting them into separate event segments. All variants follow the same training protocol and use aligned visual token budgets. We then analyze the trade-off between accuracy and compression, with the results shown in Fig. 3.

Geometry-guided evidence retention. We compare random token retention with EgoWorldMem without event segments to evaluate the effectiveness of geometry-guided retention. As shown in the first two rows of Table 2, geometry-guided retention improves accuracy on five benchmarks over random token retention, raising the average from 73.1 to 78.9, a gain of 5.8 points. In the training-free setting (Table 5), random token retention reduces average accuracy by 6.2 points relative to the full-token baseline, while geometry-guided retention limits this drop to 3.6 points. These results demonstrate that geometric redundancy estimates help allocate a limited visual token budget to more informative observations, thereby improving the VLM's video understanding performance.

Object-centric event organization. To assess the role of event organization, we compare the full EgoWorldMem with its variant without event segments (the last two rows of Table 2). Removing this organization reduces average accuracy from 80.5 to 78.9, with performance declining on all five benchmarks. VLM4D shows the largest drop of 5.3 points. These results suggest that organizing retained tokens into temporally coherent, object-specific segments helps the VLM capture brief yet critical dynamic evidence in egocentric videos.

Sensitivity to the appearance discrepancy threshold. We analyze EgoWorldMem's sensitivity to the appearance discrepancy threshold τ_Δ, which controls both dynamic-region assessment and static evidence retention. As shown in Fig. 3, the token compression rate increases with the threshold, rising from 62.34% to 77.87% as τ_Δ increases from 0.175 to 0.250. Average accuracy remains between 79.8 and 80.5 over 0.175–0.225, but falls to 79.2 at 0.250, suggesting that more aggressive selection may discard useful evidence needed to answer questions about the video. We use $\tau_\Delta = 0.225$, which achieves the highest accuracy among the tested settings with 73.96% compression. Per-benchmark results and compression rates are provided in Table 6.

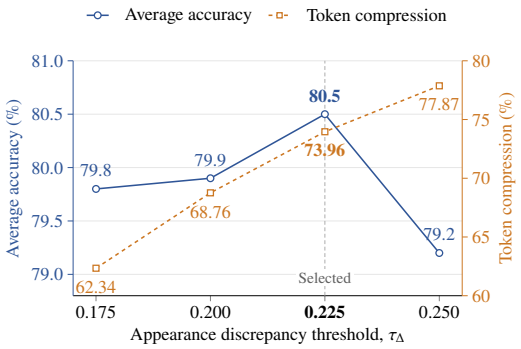


Figure 3: **Accuracy–compression trade-off.** Average accuracy and token compression versus τ_Δ. The dashed line marks the selected threshold 0.225.

4.4 Online Memory Construction and Efficiency

We first evaluate EgoWorldMem with different 3D reconstruction models. Using the streaming model CUT3R (Wang et al., 2025b), we then implement an online variant and assess its memory construction efficiency and query-time latency.

Choice of geometry backbone. We compare VGGT-Ω (Wang et al., 2026), FastVGGT (Shen et al., 2025), and CUT3R (Wang et al., 2025b) to assess EgoWorldMem's compatibility with different geometry backbones. As shown in Table 3, EgoWorldMem achieves comparable performance across the three geometry backbones, with average accuracy across the five benchmarks ranging from 79.3 to 80.5 and token compression ranging from 71.95% to 77.73%. Detailed results are provided in Table 7. In particular, the causal CUT3R variant achieves 79.8 average accuracy with 77.73% token compression, only 0.7 points below VGGT-Ω in accuracy, while reconstructing geometry in a streaming manner. These results support extending EgoWorldMem to online memory construction with limited loss in accuracy.

Table 3: **Geometry backbone comparison.** Average accuracy and token compression across five benchmarks. Detailed results are provided in Table 7.

3D Model	Stream.	Avg.	Compression(%)
VGGT-Ω	×	**80.5**	73.96
FastVGGT	×	79.3	71.95
CUT3R	✓	<u>79.8</u>	77.73

Online construction and inference efficiency. We evaluate the CUT3R-based online variant of EgoWorldMem on the EgoSchema public subset, which comprises 500 egocentric videos of 180 seconds each. We sample these videos at 1 fps and run all experiments on a single NVIDIA A800 (80 GB) GPU. As shown in Table 4, EgoWorldMem constructs memory at an average throughput of 1.283 fps, exceeding the sampled input rate, while reducing visual tokens by 83.02%. For a fair comparison, we start timing the full-token Qwen3-VL-8B baseline after its vision encoder produces visual features, while EgoWorldMem directly retrieves visual features from memory. EgoWorldMem reduces TTFT from the full-token baseline's 1595.26 ms to 679.75 ms, a reduction of 57.39%. Peak GPU memory usage increases by only 2.28 GiB, from 21.11 to 23.39 GiB. These results support the use of EgoWorldMem in egocentric assistants that continuously accumulate visual context and respond promptly to questions about past observations, with limited additional GPU memory. We provide the construction cost breakdown in Table 8.

Table 4: **Online construction and inference efficiency.** Evaluation with CUT3R on the 500-video EgoSchema public subset at 1 fps.

Method	Visual Tokens	Compression (%)	Construction Throughput (fps)	Peak Memory (GiB)	Model TTFT (ms)
Qwen3-VL-8B (full tokens)	14,913.9	0.00	–	21.11	1595.26
EgoWorldMem (online)	2,532.7	83.02	1.283	23.39	679.75

5 Conclusion

We present EgoWorldMem, a 4D world-memory framework that organizes visual history around an evolving physical world. Geometric comparisons identify repeated observations across viewpoints, while persistent object identities connect dynamic evidence into temporally and semantically coherent event segments. This organization preserves static scene context and the evolution of object activity within a compact memory for VLM inference. Experiments across five benchmarks and two VLM backbones demonstrate improved accuracy with substantially fewer visual tokens, and controlled ablations support the contributions of geometry-guided retention and event organization. EgoWorldMem also remains effective across different geometry backbones, with a streaming variant enabling incremental memory construction and faster responses once memory is ready. These findings support world-grounded memory as a foundation for egocentric assistants that retain visual context and respond efficiently and promptly to user queries.

AI USE STATEMENT

We used generative AI tools to polish the manuscript wording and create the introductory conceptual illustration. We also used Qwen3-8B to assist in constructing 1K training question–answer pairs from existing EgoIT videos (Appendix C.3). Specifically, the LLM formulated questions in natural language and generated three distractor options for each example, while answer labels were derived from estimated 3D geometry and the corresponding segmentation and tracking results.We did not use generative AI tools for other tasks requiring disclosure. We take responsibility for the final content of this work, including its claims, results, and AI-assisted artifacts.

ETHICS STATEMENT

This work studies efficient egocentric video understanding using publicly available benchmarks, including EgoSchema, QaEgo4Dv2, NextQA, NextGQA, VLM4D, and VSI-Bench, each released under its own consent and licensing terms. We propose a VLM memory and token-compression mechanism that does not involve human-subject studies or the collection of new data, and introduces no new privacy, fairness, or dual-use risks beyond those inherent to the underlying datasets, so no additional ethical review is required.

REPRODUCIBILITY STATEMENT

We describe the EgoWorldMem framework in detail in the Method section, and report the experimental setup, including benchmarks, evaluation metrics, and compared baselines, in the Experiments section. Detailed training hyperparameters are provided in the Appendix. To facilitate reproduction, we will release our code upon acceptance, together with detailed dataset download links and training scripts.

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*, 2025.

Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1560–1568, 2022.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12943, 2024.

Juntong Fang, Zequn Chen, Weiqi Zhang, Donglin Di, Xuancheng Zhang, Chengmin Yang, and Yu-Shen Liu. More: Motion-aware feed-forward 4d reconstruction transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28914–28924, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Jixuan He, Xueting Li, Chieh Hubert Lin, and Ming-Hsuan Yang. Reasmory: 3d reconstruction as explicit memory for vlms spatial reasoning. *arXiv preprint arXiv:2606.00963*, 2026.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 414–424, 2026.

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.

Han Li, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Geometry-guided 3d visual token pruning for video-language models. *arXiv preprint arXiv:2604.18260*, 2026.

Jiaao Li, Kaiyuan Li, Chen Gao, Yong Li, and Xinlei Chen. Egoprune: Efficient token pruning for egomotion video reasoning in embodied agent. *arXiv preprint arXiv:2507.15428*, 2025.

Yiren Lu, Yi Du, Disheng Liu, Yunlai Zhou, Chen Wang, and Yu Yin. Gsmem: 3d gaussian splatting as persistent spatial memory for zero-shot embodied exploration and reasoning. *arXiv preprint arXiv:2603.19137*, 2026.

Martin Q Ma, Willis Guo, Aditya Agrawal, Ankit Gupta, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Video active perception: Effective inference-time long-form video understanding with vision-language models. *arXiv preprint arXiv:2605.01662*, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

Alkesh Patel, Vibhav Chitalia, and Yinfei Yang. Advancing egocentric video question answering with multimodal large language models. *arXiv preprint arXiv:2504.04550*, 2025.

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

Qwen Team. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. Technical report, Alibaba Group, August 2026b.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

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, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024a.

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

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. *arXiv preprint arXiv:2409.01071*, 2024b.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31272–31282, 2026.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. Internvideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 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, 2025a.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17294–17303. IEEE, 2025c.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2026.

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, 2025a.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025b.

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, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

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

A METHOD AND IMPLEMENTATION DETAILS

A.1 MEMORY CONSTRUCTION HYPERPARAMETERS

For the geometry-guided redundancy criterion, we use $\tau_\Delta = 0.225$ and $\tau_\Omega = 0.3$. A visual token is associated with an object track when its source patch overlaps the track mask by at least 0.25, and an event span is retained when it contains at least 8 associated tokens. For static evidence retention, isolated components smaller than 5 patches are removed. Hyperparameters specific to outdated-geometry removal are provided with the complete update procedure in Appendix A.2.

A.2 GEOMETRIC MAP UPDATE

We update the auxiliary geometric map by removing outdated object geometry and incorporating valid observations from the current frame. The complete procedure is summarized in Algorithm 1.

Detecting outdated geometry. For each tracked object, we identify patches with at least 25% of their pixels inside its mask. These patches contribute to the object-level scores. We then examine their map coverage. If a patch has coverage below 0.3, the geometric map provides insufficient evidence to distinguish newly observed content from actual scene changes. Its appearance discrepancy is therefore set to zero before aggregation, preventing poorly observed regions from contributing spurious evidence of outdated geometry.

For objects with associated patches, we compute two object-level scores: d_o, the mean of the highest 75% of the resulting discrepancy values, and c_o, the mean of the lowest 25% of the map coverage values. The two subsets are selected independently. The object's stored geometry is removed when both $d_o \geq 0.088$ and $c_o \geq 0.2$ hold in the current frame.

Replacing stored geometry. For each object satisfying the above conditions, we remove its previously stored points from $\mathcal{P}$ and retrieve its mask in the previous frame. This mask is used to identify unassigned points that may have been stored while the object was static and not yet tracked. To remove such points, we project the remaining geometric map into the previous camera view and select unassigned points ($o_j = -1$) inside the mask that satisfy

$$z_j \leq z_{\min} + \max(0.02,\, 0.08 z_{\min}), \tag{6}$$

where $z_{\min}$ is the smallest positive depth of the projected map points at the same pixel. This condition restricts removal to the frontmost surface and its vicinity, preserving deeper background geometry.

After removal, valid current points are inserted with their colors, reconstruction confidences, timestamps, and object identities. Object-wise voxel merging is then applied when its trigger threshold is exceeded, as detailed in Appendix A.3.

A.3 MAP BUDGET AND VOXEL MERGING

To keep the geometric map bounded, we cap the number of geometric entries of each video at $N_{\max} = 2 \times 10^6$. Merging is triggered only when a video's geometric entry count exceeds $\rho N_{\max}$ with $\rho = 1.2$; below this threshold the point cloud is kept lossless. When triggered, the occupied volume is partitioned into a voxel grid whose resolution is chosen so that the number of voxels matches the budget, and all entries falling into the same voxel are consolidated into a single representative entry, averaging positions, colors, and confidences and keeping the latest timestamp. The budget is distributed across objects in proportion to each object's share of the points, so a small dynamic object retains its own allocation and is not absorbed by the dominant background.

A.4 VISUAL MEMORY SIZE CONTROL

The geometric map budget in Appendix A.3 controls the size of $\mathcal{P}_t$, while the visual memory $\mathcal{M}_t$ stores retained static evidence and dynamic event tokens. Its size can be adjusted through the appearance discrepancy threshold τ_Δ and the coverage threshold τ_Ω. These thresholds jointly determine how incoming observations are classified and retained, thereby regulating the amount of evidence

Algorithm 1: Geometric Map Update

Input: Geometric map $\mathcal{P}$, current reconstruction, tracked object masks, previous reconstruction and masks
Output: Updated geometric map $\mathcal{P}$

```
Compute patch appearance discrepancy and map coverage
Initialize the previous-mask collection S ← ∅
foreach tracked object o do
    Collect patches with object-mask overlap ≥ 0.25
    if no patches are associated then
        continue
    Set discrepancies to zero where map coverage < 0.3
    d_o ← mean of the highest 75% of discrepancies
    c_o ← mean of the lowest 25% of coverage values
    if d_o ≥ 0.088 and c_o ≥ 0.2 then
        Remove points labeled o from P
        Retrieve the previous mask; backward-track if unavailable
        Add the previous mask to S if available
if the previous reconstruction is available and S ≠ ∅ then
    Project the remaining P into the previous view
    Compute the frontmost depth z_min at each pixel
    Remove unassigned points inside the union of masks in S that satisfy the surface-depth condition
Insert valid current points with colors, confidences, timestamps, and object labels
Apply object-wise voxel merging if its trigger is exceeded
return P
```

added to memory. The effect of τ_Δ on the accuracy–compression trade-off is evaluated in Fig. 3. Such threshold-based selection reduces memory growth, but does not by itself impose a fixed upper bound on accumulated visual tokens.

For longer streams, the object-centric representation also supports an optional token-merging extension. When the visual-token count exceeds a target budget, temporally or spatially nearby tokens with similar features can be merged within the same object's event segment. The resulting representative tokens retain the object association and representative source timestamps and patch positions for memory readout. Keeping event boundaries and temporal ordering helps limit the mixing of evidence from different object states. Static background tokens can be compressed separately within spatial neighborhoods, allowing both components of $\mathcal{M}_t$ to be reduced.

Threshold adjustment and token merging provide complementary ways to control incoming evidence and compress accumulated history. Enforcing a strict global budget would additionally require a policy for pruning event or background groups when within-group merging is insufficient, since the number of distinct objects and events can continue to grow. We describe this budget-aware extension as a possible approach to longer streams; its accuracy–memory trade-off requires further evaluation.

B Experimental Settings

B.1 Training Details

We fully fine-tune Qwen3-VL-8B and LLaVA-NeXT-Video using the same training configuration on eight NVIDIA A800 GPUs with 80 GB of memory each. Both models are trained for two epochs with a learning rate of 10^{-5} and a cosine learning-rate schedule, with warmup over the first 10% of training steps. We use a per-device batch size of 1 and accumulate gradients over 8 steps, yielding an effective global batch size of 64. Training uses BF16 mixed precision, FlashAttention-2, and DeepSpeed ZeRO-2. We sample the EgoIT videos used for VLM4D-style QA at 4 fps and all other training videos at 1 fps. The minimum and maximum pixel budgets per frame are set to $20{,}480$ and $184{,}320$, respectively.

B.2 Proprietary API-Based Models

For GPT-4o, we use the official VLM4D Overall score of 57.5 under chain-of-thought prompting (Zhou et al., 2025). The EgoSchema and NextQA scores are direct baselines reproduced by Video Active Perception (Ma et al., 2026): GPT-4o receives frames uniformly sampled at 1 fps, amounting to 180 frames for each three-minute EgoSchema video and approximately 44 frames per NextQA video. EgoSchema uses chain-of-thought prompting, whereas NextQA requests only the answer index. These experiments report 67.8 on the EgoSchema full test set and 71.2 on the NextQA test set. For QaEgo4Dv2, we report the full-video CloseQA result of 26.28 ± 8.1 (Patel et al., 2025), where 32 frames are uniformly sampled over the full video, resized to 336×336, and passed directly to the model with a zero-shot prompt that requests the selected answer without an explanation. The cited QaEgo4Dv2 results and our evaluation use the complete test set and original answer choices, with accuracy averaged over five random seeds for answer-choice shuffling. No direct GPT-4o result is available on the complete NextGQA test set.

For Gemini-1.5 Pro, the EgoSchema score of 72.2 is the official zero-shot chain-of-thought result from the Gemini 1.5 technical report (Team et al., 2024); the exact frame count for this entry is not disclosed. Video Active Perception independently reproduces 65.8 by uploading the native video to the Gemini API under its 1-fps-equivalent baseline setting (Ma et al., 2026). On QaEgo4Dv2, Gemini-1.5 Pro obtains 50.76 ± 1.4 under the same 32-frame full-video CloseQA protocol described above (Patel et al., 2025). On NextQA, it obtains 70.4 when the native video is uploaded to the Gemini API; the source categorizes this setting as 1 fps, or approximately 44 frames per video, and uses a direct-answer prompt (Ma et al., 2026). No standard Gemini-1.5 Pro result is available on VLM4D or the complete NextGQA test set. We exclude results that use external captioners, adaptive frame selection, retrieval, agents, model ensembles, or fine-tuning.

B.3 Open-Source and Memory-Augmented VLMs

For the memory-augmented baselines evaluated in Table 1, we tune each method's compression parameters to match EgoWorldMem's retained visual-token budget on each benchmark, targeting a deviation within $\pm 10\%$ where feasible. The configurations and method-specific constraints are detailed below. Percentage deviations are measured relative to the corresponding EgoWorldMem token budget.

LiveVLM. LiveVLM uses a training-free streaming KV cache with LLaVA-OneVision-7B. Its `cache_size` parameter directly caps the number of visual KV tokens. On EgoSchema, setting this parameter to the target budget yields 3,965 retained tokens, matching EgoWorldMem's budget exactly. Across all five benchmarks, the retained-token counts remain within $\pm 3\%$ of the corresponding targets.

Flash-VStream. Flash-VStream uses temporal and spatial memory banks with Qwen2-VL-7B. We jointly tune the memory-bank capacities in `memory_dict`, the frame limit `max_frames`, and the pixel budget `max_pixels`. On EgoSchema, a temporal/spatial memory setting of $60/30$, a frame limit of 110, and a pixel budget of 150,528 yield 4,219 retained tokens, approximately 6% above the target. The retained-token counts fall within $\pm 10\%$ of the targets on all five benchmarks.

FluxMem. FluxMem uses training-free hierarchical memory with Qwen2.5-VL-7B. Because it does not expose a direct token-budget parameter, we tune `medium_frames`, `max_pixels`, and `max_num_frames` to approximate the target budget. On EgoSchema, setting `medium_frames` to 48 and the pixel budget to 87,808 yields 4,198 retained tokens, approximately 6% above the target. All five benchmarks meet the $\pm 10\%$ matching criterion.

StreamForest. StreamForest uses PEMF event memory with LLaVA-OneVision-7B. We control its retained-token count through `max_frames_num`. On EgoSchema, a limit of 30 frames yields 4,065 tokens, approximately 3% above the target. Its event-memory representation imposes a discrete lower bound on the token count: 1,113 tokens for frame limits of at most 4, and 1,625 tokens at 8 frames. Consequently, the smaller target budgets on NExT-QA, NExT-GQA, and VLM4D cannot be matched; we report these results at the minimum attainable budget.

C Evaluation Metrics and Benchmarks

C.1 Metric Aggregation

We report multiple-choice QA accuracy on each of the five main benchmarks. Overall accuracy is their unweighted arithmetic mean. Visual-token compression is computed from pooled token counts. Let N_i^{full} and N_i^{keep} denote the numbers of visual tokens before and after compaction for evaluation input i, respectively. The reported compression percentage measures token reduction:

$$\text{Compression}(\%) = \left(1 - \frac{\sum_{i \in \mathcal{E}} N_i^{\text{keep}}}{\sum_{i \in \mathcal{E}} N_i^{\text{full}}}\right) \times 100\%, \tag{7}$$

where $\mathcal{E}$ contains the evaluation inputs included in the reported result. Per-benchmark compression uses inputs from that benchmark; overall compression pools token counts across all five benchmarks. Thus, overall compression is weighted by the original visual-token counts, rather than being the arithmetic mean of per-benchmark percentages. The reported reductions of 73.96% and 78.65% correspond to retaining 26.04% and 21.35% of the original visual tokens for Qwen3-VL and LLaVA-NeXT-Video, respectively.

C.2 Benchmarks

VLM4D. VLM4D (Zhou et al., 2025) evaluates spatiotemporal reasoning, including translational and rotational motion, perspective awareness, and motion continuity. It combines real egocentric videos from Ego4D, real exocentric videos from DAVIS and YouTube-VOS, and synthetic videos. We report its overall multiple-choice QA accuracy.

EgoSchema. EgoSchema (Mangalam et al., 2023) evaluates understanding of three-minute egocentric clips drawn from Ego4D. Each question has five answer choices and is designed to require extended temporal context. It complements the motion-focused evaluation of VLM4D with reasoning over sustained activities. The efficiency experiment uses its 500-video public subset, as specified in Section 4.4.

QaEgo4Dv2. QaEgo4Dv2 (Bärmann & Waibel, 2022; Di & Xie, 2024; Patel et al., 2025) is a refined version of QaEgo4D for question answering about past observations in egocentric videos. We use the full-video CloseQA setting with the original answer choices and report accuracy on the complete test set, averaged over five answer-choice shuffling seeds, following the protocol described in Appendix B.2.

NExT-QA. NExT-QA (Xiao et al., 2021) evaluates causal and temporal action reasoning together with descriptive understanding of everyday video content. We use its multiple-choice QA setting to assess how the retained evidence supports reasoning about actions and their temporal relationships.

NExT-GQA. NExT-GQA (Xiao et al., 2024) adds temporal grounding annotations to subsets of the NExT-QA validation and test QA pairs, excluding questions that depend on global video content and descriptive questions. We evaluate its multiple-choice QA component and report answer accuracy. Because NExT-QA and NExT-GQA share evaluation examples, their results provide correlated evidence. We report both separately and use the five-benchmark mean as a descriptive aggregate, without treating the five results as independent evaluations.

VSI-Bench (additional spatial evaluation). VSI-Bench (Yang et al., 2025a) evaluates spatial reasoning from video through numerical and multiple-choice questions. Its metrics are mean relative accuracy for numerical answers and accuracy for multiple-choice answers. We use it to study spatial position embedding in Appendix E; these results are reported separately and are excluded from the five-benchmark aggregates.

C.3 Construction of EgoIT Training QA Pairs

We construct 1K training question–answer pairs from EgoIT videos sampled at 4 fps. For each video, we use VGGT (Wang et al., 2025a) to estimate camera parameters and dense 3D point maps,

and apply SAM2 (Ravi et al., 2025) to segment and track objects across frames. Associating the tracked object masks with the estimated 3D points provides object-level geometric observations over time, linking spatial information with persistent object identities.

Following VLM4D's question formats (Zhou et al., 2025), we construct motion-related questions covering translational motion, rotational motion, spatiotemporal counting, and false positives. These questions concern object movement directions, rotation directions, the number of distinct objects satisfying a specified motion condition, and absent objects or motions that do not occur. We derive the answer labels from the estimated 3D information and the corresponding segmentation and tracking results, interpreting directional descriptions from the viewpoint specified in each question.

We use Qwen3-8B to organize the questions into natural language and generate three distractor options for each derived answer. Each example therefore consists of a question, four answer choices, and the corresponding answer label. The resulting 1K QA pairs are included in the training set for fine-tuning both VLM backbones.

D Additional Experimental Results

D.1 Controlled Fine-Tuning Comparison

Full-token fine-tuning on eight NVIDIA A800 GPUs (80 GB each) encountered out-of-memory errors with DeepSpeed ZeRO-3. Matching EgoWorldMem's token budget by uniformly subsampling whole frames would substantially coarsen temporal sampling. For example, assuming comparable token counts per frame, the 73.96% token reduction reported for EgoWorldMem with Qwen3-VL would reduce a 1-fps input to approximately 0.26 fps, or one frame every $1/(1-0.7396) \approx 3.84$ seconds. Such sparse sampling misses brief interactions and other task-relevant temporal evidence, thereby impairing model performance. We therefore use random token retention to match the visual-token budget while keeping the source-frame sampling schedule unchanged.

To assess whether the performance gains can be explained by task-specific training data and fine-tuning alone, we fine-tune this baseline using the same Qwen3-VL-8B backbone, training data, and fine-tuning protocol as EgoWorldMem. Specifically, we retain the same input frame sampling rates (1 fps for NExT-QA and QaEgo4Dv2, and 4 fps for EgoIT) and randomly select visual tokens from the resulting sequence to match EgoWorldMem's per-video token budget.

As shown in Table 2, the fine-tuned random-retention baseline achieves an average accuracy of 73.1, compared with 80.5 for EgoWorldMem. The 7.4-point advantage persists when training data, fine-tuning protocol, and visual-token budget are held fixed. This comparison shows that applying the same task-specific training to random token retention does not reproduce EgoWorldMem's performance, supporting the value of its memory design beyond access to additional training data.

D.2 Training-Free Evidence Selection

In the training-free setting, we preserve the original token-sequence structure and compare geometry-guided retention with random token retention under the same token budget. As shown in Table 5, geometry-guided retention achieves an average accuracy of 66.9, compared with 64.3 for random retention. It remains 3.6 points below the full-token baseline of 70.5. These results show that geometry-guided selection retains more useful evidence than random retention even without model adaptation.

Table 5: **Training-free evidence selection on Qwen3-VL-8B.** Comparing geometry-guided retention with random token retention in the training-free setting. The full-token baseline provides an uncompressed reference. Avg. denotes mean accuracy across the five benchmarks.

Memory configuration	Accuracy (%)					
	VLM4D	EgoSchema	QA-Ego4D	NextQA	NextGQA	Avg.
Full-token baseline	56.0	71.8	63.9	81.9	79.1	70.5
Random token retention	49.0	65.8	56.3	76.7	73.9	64.3
Geometry-guided retention	53.3	67.9	57.9	78.4	77.1	66.9

D.3 Sensitivity to the Appearance Discrepancy Threshold

We report per-benchmark accuracy and visual-token compression for different values of τ_Δ in Table 6, complementing the trade-off analysis in Fig. 3.

Table 6: **Per-benchmark results for the appearance discrepancy threshold ablation.** Positional embeddings are disabled in all variants. Avg. denotes the average accuracy across the five benchmarks; Compression reports the visual-token reduction. Bold and underlined values denote the best and second-best results, respectively.

τ_Δ	VLM4D	EgoSchema	QA-Ego4D	NextQA	NextGQA	Avg.	Compression
0.175	74.2	83.1	65.9	**88.7**	**87.1**	79.8	62.34%
0.200	<u>74.3</u>	**84.2**	65.7	88.3	<u>86.9</u>	<u>79.9</u>	68.76%
0.225	**74.9**	83.6	**68.5**	<u>88.6</u>	**87.1**	**80.5**	73.96%
0.250	73.1	<u>84.0</u>	<u>66.1</u>	87.4	85.4	79.2	77.87%

D.4 Effect of Geometry Backbone

Table 7 provides the per-benchmark results for the three geometry backbones compared in Table 3.

Table 7: **Per-benchmark results for the 3D reconstruction backbone ablation.** Avg. denotes the average accuracy across the five benchmarks. Bold and underlined values denote the best and second-best results, respectively.

3D Model	VLM4D	EgoSchema	QA-Ego4D	NextQA	NextGQA	Avg.
VGGT-Ω (Wang et al., 2026)	**74.9**	<u>83.6</u>	**68.5**	**88.6**	**87.1**	**80.5**
FastVGGT (Shen et al., 2025)	**74.9**	**84.3**	63.1	87.9	86.2	79.3
CUT3R (Wang et al., 2025b)	<u>74.6</u>	<u>83.6</u>	<u>66.5</u>	<u>88.0</u>	<u>86.4</u>	<u>79.8</u>

D.5 Memory Construction Cost

Table 8: **Memory construction cost of online EgoWorldMem.** EgoSchema public subset, 1 fps, CUT3R, one A800 (80 GB) GPU. Compaction includes vision encoding. Total is derived from throughput; Others is the residual after subtracting the three timed stages.

3D reconstruction (ms/frame)	SAM2 (ms/frame)	Compaction (ms/frame)	Others (ms/frame)	Total (ms/frame)	Overall throughput (fps)
49.71	341.05	379.74	8.92	779.42	1.283

We report model TTFT, measured from the start of inference on visual token features to the generation of the first output token. For a fair comparison, timing for the full-token Qwen3-VL-8B baseline begins after vision encoding. At query time, EgoWorldMem directly consumes visual tokens precomputed during memory construction.

E SPATIAL REASONING ON VSI-BENCH

E.1 EXPERIMENTAL SETUP

We evaluate EgoWorldMem on VSI-Bench (Yang et al., 2025a) to explore the effect of spatial position embedding (SPE) on spatial reasoning from video. For this experiment, we fine-tune the model on VSI-Train-10K and VSI-590K for 6 epochs and set the appearance discrepancy threshold to $\tau_\Delta = 0.23$. The compact memory retains approximately 29.26% of the original visual tokens, corresponding to a token compression rate of 70.74%. We compare variants with and without SPE to assess the benefit of explicitly encoding the memory's world-space grounding. SPE is an optional extension and is disabled in the main video-understanding experiments.

E.2 SPATIAL POSITION EMBEDDING

Each retained token already carries a 4D anchor (τ_i, x_i, y_i, z_i) through world memory grounding (Section 3.4). SPE converts this anchor into an additive embedding, allowing the VLM to access the token's location in a shared world coordinate frame together with its observation time. In the implementation, τ_i is represented by the source frame index. For each scalar coordinate q, we use $L = 8$ frequency bands to encode position at multiple scales and concatenate the encodings of all four coordinates:

$$\gamma(q) = \left[q, \left(\sin(2^k q)\right)_{k=0}^{L-1}, \left(\cos(2^k q)\right)_{k=0}^{L-1}\right], \qquad \mathbf{e}_i = \left[\gamma(\tau_i); \gamma(x_i); \gamma(y_i); \gamma(z_i)\right]. \tag{8}$$

Each coordinate contributes its raw value and L sine–cosine pairs, giving a concatenated dimension of $4(2L + 1) = 68$. A two-layer projection maps this encoding to the VLM's input embedding dimension and adds it to the visual token feature:

$$\widetilde{f}_i = f_i + W_2\ \mathrm{GELU}(\mathrm{LN}(W_1 \mathbf{e}_i + b_1)) + b_2, \tag{9}$$

where W_1, W_2, b_1, b_2 are learnable parameters and LN denotes layer normalization. This embedding is applied to both retained static and dynamic visual tokens before the language model, while the existing M-RoPE position IDs are preserved. Thus, SPE supplies explicit world-space and temporal information without increasing the memory token count.

E.3 RESULTS AND ANALYSIS

EgoWorldMem with SPE achieves an overall score of 68.6 on VSI-Bench, as reported in Table 9. The comparison in Table 10 isolates the effect of SPE: enabling it raises the overall score from 67.09 to 68.56, a gain of 1.47 points. The largest improvements occur in route planning ($+4.64$ points), absolute distance ($+3.91$), and relative direction ($+3.24$).

These gains are consistent with the spatial cues supplied by SPE. Encoding observations in a shared world coordinate frame can make their spatial relationships more accessible to the VLM across changes in viewpoint, helping it estimate distances and directions and relate successive views when planning routes. The benefit varies across subtasks: improvements in room size, relative distance, and appearance order are smaller, while object counting and object size decrease by 0.32 and 1.03 points, respectively. This pattern suggests that explicit coordinate embeddings are particularly useful for reasoning about spatial relationships, although they do not improve every spatial task.

On the video-understanding tasks used in the main experiments, enabling SPE yields slightly worse overall performance than omitting it. We therefore disable SPE in all main experiments and explore it here as an extension for spatial reasoning. The main model still uses geometry to estimate redundancy and construct the world-grounded memory, while SPE additionally encodes the tokens' 4D anchors into their visual features. Taken together, these observations suggest that the benefit of explicit coordinate embeddings depends on the spatial demands of the task.

Table 9: **Evaluations on VSI-Bench (Yang et al., 2025a) for 3D spatial reasoning tasks.** We compare against proprietary models, open-sourced VLMs, and spatial-enhanced models specifically designed for spatial reasoning. Bold and underlined values denote the best and second-best results, respectively. The Improve row reports the gain of EgoWorldMem over its Qwen3-VL-8B backbone.

Methods	Rank	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	3	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	2	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	1	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced VLMs										
LongVA-7B	12	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
InternVL2-8B	8	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	6	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LongVILA-8B	14	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-8B	13	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	11	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
Qwen2.5VL-7B	9	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
Qwen2.5VL-72B	5	37.0	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9
LLaVA-OneVision-7B	10	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	4	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B	7	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	3	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Qwen3-VL-8B	2	59.4	<u>70.6</u>	50.9	<u>75.1</u>	60.5	59.9	51.2	33.0	<u>74.4</u>
Qwen3-VL-32B	1	<u>63.8</u>	**74.2**	**56.5**	**78.0**	<u>69.2</u>	57.5	56.1	44.3	74.3
Spatial-Enhanced Models										
VG-LLM-4B	5	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
VG-LLM-8B	3	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
Spatial-MLLM-4B	4	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VLM-3R-7B	2	60.9	70.2	49.4	69.2	67.1	**65.4**	<u>80.5</u>	<u>45.4</u>	40.1
EgoWorldMem-8B (Ours)	1	**68.6**	69.7	<u>53.5</u>	72.6	**73.8**	<u>63.9</u>	**85.9**	**45.9**	**83.0**
Improve ↑	-	+9.2	-0.9	+2.6	-2.5	+13.3	+4.0	+34.7	+12.9	+8.6

Table 10: **Effect of spatial position embedding on VSI-Bench.** All scores are reported on a 0–100 scale. Overall uses the aggregate score reported by the evaluator. Bold values denote the better result in each column.

Variant	Overall	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
w/o SPE	67.09	**70.04**	49.63	**73.66**	73.26	63.52	82.71	41.24	82.69
w/ SPE	**68.56**	69.72	**53.54**	72.63	**73.82**	**63.94**	**85.95**	**45.88**	**83.01**