EGOPROBE: BENCHMARKING LONG-RANGE AND EXACT MEMORY IN STREAMING EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

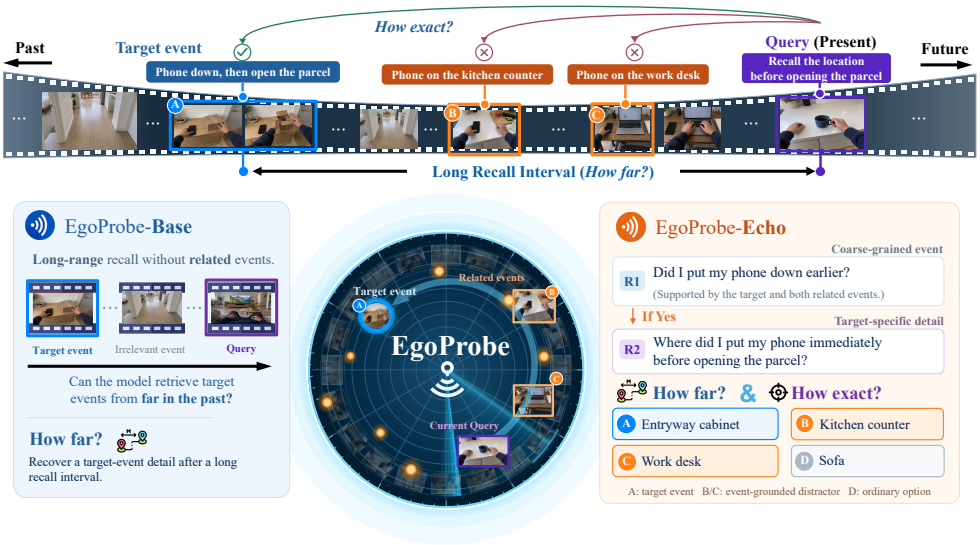


Figure 1: **Overview of EgoProbe.** EgoProbe-Base probes long-range recall without intervening related events. EgoProbe-Echo further uses coarse-to-fine questions and event-grounded distractors to probe recall of target-event details among related experiences.

ABSTRACT

Egocentric video provides a continuous first-person record of everyday activities, making it a natural foundation for wearable AI assistants. Unlike offline video understanding, streaming systems must process observations as they arrive, while queries may refer to experiences from the distant past. Recent work has begun to explore whether information can be recalled across long temporal gaps. Yet remembering how far is only part of the challenge: recurring objects, actions, and locations also require models to remember exactly which detail belongs to which event. We introduce **EgoProbe**, a benchmark that evaluates streaming egocentric memory along these two complementary dimensions: long-range retention, i.e., *how far*, and event-specific recall under interference, i.e., *how exact*. Built on long-form egocentric videos, EgoProbe uses an event-grounded data engine to extend recall intervals, identify related events, and construct event-grounded distractors, followed by human verification. It contains two subsets: **EgoProbe-Base**, which evaluates long-range recall without intervening related events, and **EgoProbe-Echo**, which further challenges models with similar intervening events and event-grounded distractors. EgoProbe-Echo additionally adopts coarse-to-fine multi-turn probing: the first turn evaluates coarse-grained event recall, while the second tests whether the model can retrieve the target-specific detail associated with the designated event. Extensive evaluation across diverse models reveals substantial limitations in long-range recall, with further challenges in recovering target-event details among related experiences. Even when models correctly answer coarse-grained questions, they often fail to recall the specific details of the designated event. EgoProbe exposes this underexplored failure mode and provides a testbed for advancing long-range and exact memory in streaming egocentric video understanding. Code and data will be released.

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.

1 INTRODUCTION

Wearable AI assistants, e.g., always-on intelligent glasses, are expected to perceive continuously as users observe and interact with the world Grauman et al. (2022); Damen et al. (2020). In such real-world settings, useful information is scattered across a long, evolving visual stream rather than confined to a short and self-contained video clip. A query raised at the current moment may depend on an object seen, an action performed, or an interaction that occurred minutes or even hours earlier Bärmann & Waibel (2022). Supporting such assistance therefore requires models to move beyond local perception to long-range memory for egocentric video streams.

Despite growing interest in streaming video understanding, existing benchmarks Niu et al. (2025); Lin et al. (2026); Fu et al. (2025) remain limited in three important aspects. First, many benchmarks Li et al. (2024b); Fu et al. (2025) are designed for generic video rather than the specific demands of egocentric streams, which present distinctive challenges including restricted fields of view, substantial ego-motion, and frequent interactions with surrounding. Second, a streaming protocol alone does not necessarily imply long-range memory dependence: many queries may still be answered from recent observations, without requiring information to be retained across long recall intervals Shen et al. (2026). Third, recent work Forte et al. (2026) has begun to study long-range memory in egocentric streaming settings, largely focusing on how far information can be recalled. However, temporal distance alone does not fully characterize reliable memory. In everyday egocentric streams, similar events may recur, requiring models not only to retain information over time but also to associate each recalled detail with the correct event.

These limitations naturally motivate the design of **EgoProbe** (Fig. 1), which is designed to 1) reflect the distinctive characteristics of *egocentric streams*; 2) explicitly require *long-range* memory; and 3) also enable the evaluation of *how exact* the models can recall details of the target event.

To achieve these goals, we build EgoProbe on two representative long-form egocentric video sources, Ego4D Grauman et al. (2022) and EgoLife Yang et al. (2025), which naturally provide the first-person, continuous visual streams required for egocentric evaluation and form a strong basis for studying long-range memory. Their associated episodic QA annotations, however, were not specifically designed to assess either long-range memory or precise recall of target-event details in the presence of related events. We therefore develop an event-grounded data engine to transform these questions into controlled probes of long-range memory and event-specific recall. For *long-range* memory, we use delayed query to control the recall interval between the target-event evidence and the query. Since additional related events before the delayed query may make the original question ambiguous, the engine further checks and, when necessary, disambiguates the question's reference to the target event to ensure that each extended question remains uniquely and correctly grounded in the video. For *exact* memory, we exploit recurring event patterns in egocentric streams to construct challenging distractors from related events that share similar semantics but differ in event-specific details. Rather than using arbitrary incorrect options, these distractors are grounded in actual observations from the same stream, allowing us to test whether a model recalls the correct detail of the target event without confusing it with details from related events. We further extend this evaluation into coarse-to-fine two-turn dialogues: the first turn asks about shared coarse event content, while the second queries the target-specific detail. This design enables EgoProbe to distinguish coarse-grained event recall from target-specific detail recall.

EgoProbe consists of two complementary subsets, both spanning various duration of recall intervals. **EgoProbe-Base** contains 1,137 single-turn questions without intervening related events; **EgoProbe-Echo** contains 1,229 single-turn questions with contrastive related events, together with 1,229 derived coarse-to-fine two-turn dialogues, enabling evaluation of event-specific memory under interference. All final QA pairs and their associated event evidence undergo human verification and refinement. Together, the two subsets support complementary analysis of long-range retention and event-specific recall in streaming egocentric video.

We evaluate a diverse range of advanced MLLMs across different model families. Our results reveal substantial gaps in both long-range memory and event-specific detail recall. On EgoProbe-Base, models struggle to recall distant target events even without intervening related experiences, with particularly low accuracy over ultra-long intervals. On EgoProbe-Echo, models frequently select details belonging to other related events, while successful coarse-grained responses do not ensure

correct recall of target-specific details. These findings highlight the remaining challenges in retaining information over extended streams and recalling the correct details of a designated experience.

Our contributions are threefold: (1) We introduce **EgoProbe**, a benchmark for long-range and event-specific memory in streaming egocentric video, combining recall-interval evaluation with event-grounded distractors and coarse-to-fine two-turn probing. (2) We develop an event-grounded data engine and curate 2,366 independent QA pairs with 1,229 derived two-turn dialogues, supported by human verification of QA quality and associated event evidence. (3) We extensively evaluate diverse model families, revealing limitations in long-range recall, errors associated with other related experiences, and gaps between coarse-grained responses and target-specific detail recall.

2 Related Work

2.1 Egocentric Video Understanding

Egocentric video understanding has attracted growing attention, giving rise to numerous datasets and benchmarks. Representative datasets such as EPIC-KITCHENS Damen et al. (2020), Ego4D Grauman et al. (2022), HD-EPIC Perrett et al. (2025), and Ego-Exo4D Grauman et al. (2024) support the study of everyday activities, human–object interactions, and ego-exo relations, while benchmarks such as EgoSchema Mangalam et al. (2023), EgoTempo Plizzari et al. (2025), and EgoCross Li et al. (2026b) explore egocentric video understanding from different perspectives. Beyond these tasks, growing attention has been devoted to understanding and recalling past experiences over extended periods. QaEgo4D Bärmann & Waibel (2022), derived from Ego4D Grauman et al. (2022), investigates episodic-memory question answering, while EgoLife Yang et al. (2025) explores long-context understanding of prolonged egocentric recordings for everyday life assistance. While these datasets provide valuable foundations for studying memory over extended first-person observations, their use in streaming settings remains comparatively underexplored, despite the growing need for always-on wearable assistants that continuously process observations and maintain access to past information. Building on Ego4D Grauman et al. (2022) and EgoLife Yang et al. (2025), EgoProbe investigates long-range and event-specific memory in this streaming setting.

2.2 Offline and Long Video Understanding

With the rapid development of multimodal large language models (MLLMs), video understanding has advanced significantly. Models such as LLaVA Li et al. (2024a), InternVL Wang et al. (2025), and Qwen Team et al. (2025), alongside proprietary models including Gemini Team et al. (2023) and GPT Achiam et al. (2023), have demonstrated strong capabilities across diverse video understanding tasks. Building upon these general-purpose MLLMs, recent approaches have further explored long-context video modeling. For example, LongVA Zhang et al. (2024) transfers long-context capabilities from language to vision, LongVU Shen et al. (2024) employs adaptive spatiotemporal compression, and VideoChat-Flash Li et al. (2026a) introduces hierarchical compression for efficient long-video understanding. Despite these advances, offline video understanding methods typically assume that the entire video is available at query time, whereas streaming systems must process observations as they arrive and maintain memory for future queries.

2.3 Streaming Video Understanding and Benchmarks

Recent streaming video approaches explore online interaction, memory management, and structured event representations. VideoLLM-online Chen et al. (2024) develop streaming video dialogue systems, while Flash-VStream Zhang et al. (2025) investigate memory mechanisms for continuous observation. StreamingVLM Xu et al. (2026) and HERMES Zhang et al. (2026) explore KV-cache management and hierarchical memory, respectively. StreamForest Zeng et al. (2026), OASIS Liang et al. (2026), and EventMemAgent Wen et al. (2026) focus on structured event memory and retrieval, while Video Streaming Thinking Guan et al. (2026) and JoyAI-VL-Interaction Yao et al. (2026) interleaves reasoning with incoming observations.

Existing benchmarks, including StreamingBench Lin et al. (2026), OVBench Huang et al. (2025), OVO-Bench Niu et al. (2025), and RIVER Shi et al. (2026), evaluate online perception, contextual understanding, retrospective reasoning, and anticipation. More recent benchmarks explicitly inves-

Table 1: Comparison with related benchmarks. △ indicates partially egocentric video coverage. Event-grounded distractors are options explicitly linked to evidence from other related events.

Benchmark	Egocentric	Streaming	Memory Recall	Event-grounded Distractors	Multi-turn Dialogue
EgoTempo Plizzari et al. (2025)	✓	✗	✗	✗	✗
EgoLife / EgoLifeQA Yang et al. (2025)	✓	✗	✓	✗	✗
OVO-Bench Niu et al. (2025)	△	✓	✓	✗	✗
RIVER Shi et al. (2026)	△	✓	✓	✗	✗
StreamingBench Lin et al. (2026)	✗	✓	✓	✗	✓
EgoStream Forte et al. (2026)	✓	✓	✓	✗	✗
EgoProbe (Ours)	✓	✓	✓	✓	✓

tigate streaming memory and interaction. EGOSTREAM Forte et al. (2026) evaluates episodic memory across recall intervals and memory dimensions through answer-validity windows. EgoSAT Lei et al. (2026) covers retrospective, ongoing, and prospective interaction understanding, while S-EMBER Wang et al. (2026) evaluates grounded episodic retrieval with supporting temporal evidence. Complementing these benchmarks, EgoProbe jointly evaluates long-range retention and exact event-specific recall using distractors grounded in related events. Its coarse-to-fine dialogues further distinguish coarse event recall from target-specific detail retrieval. Table 1 summarizes the key differences between EgoProbe and related benchmarks.

3 EGOPROBE BENCHMARK

3.1 TASK FORMULATION AND CORE DESIGN PRINCIPLES

EgoProbe evaluates long-range memory and event-specific recall through retrospective question answering in streaming egocentric video. Given a streaming egocentric video V, a model processes incoming observations and continuously updates its memory without knowing future queries. At time t_q, the model receives a question q about a past target event $e^\star$ and must retrieve the requested detail from its accumulated memory.

Enabling How Far: Long-range memory evaluation. To explicitly evaluate long-range memory, we define a controllable *recall interval*, $\Delta t = t_q - t_e^\star$, where $t_e^\star$ denotes the end of the target-event evidence. By adjusting the query arrival time, eligible questions can be extended from short to substantially longer recall intervals, requiring models to retrieve information observed much earlier in the stream. However, extending the recall interval may introduce additional events that make the original question ambiguous with respect to its target event. We therefore apply *query disambiguation* to ensure that each extended question uniquely identifies its target event while preserving the original answer.

Enabling How Exact: Event-specific recall evaluation. To evaluate whether models can associate recalled details with the target event, we define *related events* as distinct occurrences that share a question-relevant semantic pattern with the target event, such as a common action, object, or activity. We use alternative answers supported by these related events as *event-grounded distractors*: details that are valid for another experience but incorrect for the target event. These distractors require models to recall the requested target-event detail without confusing it with details from similar experiences in the stream. As illustrated in Fig. 1, a user put down a phone in an entryway cabinet before opening a parcel, and later place it on a kitchen counter and a work desk. For the query "Where did I put my phone before opening the parcel?", only the first event is the target event, while the later locations serve as event-grounded distractors. Correct recall therefore requires associating "entryway cabinet" with the designated event rather than recalling another valid phone location from the stream. To further distinguish different levels of memory, we introduce *coarse-to-fine two-turn probing*. The first round (R1) asks a yes/no question about the shared event pattern, e.g., "Did I put my phone down?" An affirmative response triggers the second round (R2), which asks for the detail associated with the designated target event, e.g., "Where did I put it before opening the parcel?"

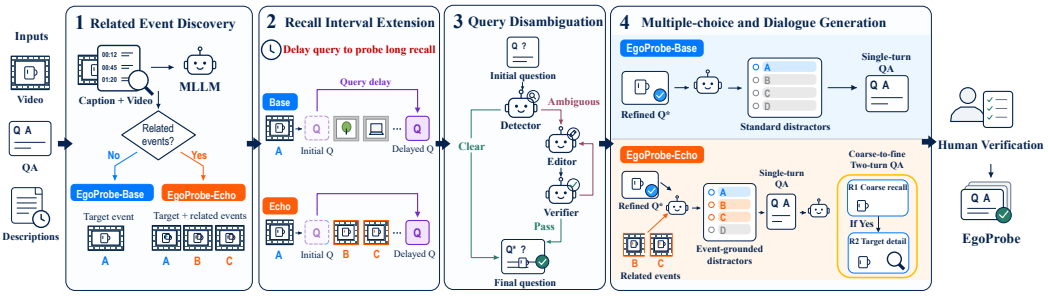


Figure 2: **EgoProbe data engine.** Four automatic stages transform source QA into streaming memory probes: related-event discovery, recall interval extension, query disambiguation, and multiple-choice and dialogue generation. The resulting QA items are then reviewed by human annotators.

3.2 Data Engine for Benchmark Construction

Overview. Our data engine starts from egocentric videos with dense temporal descriptions and existing QA annotations, using QaEgo4D Bärmann & Waibel (2022), which provides QA annotations for Ego4D Grauman et al. (2022) videos, and EgoLife Yang et al. (2025)as source datasets. Given a video, a target-event QA pair (q, a), timestamped dense captions, and its original evidence interval, we construct two complementary subsets. **EgoProbe-Base** contains samples without related events within the recall interval, whereas **EgoProbe-Echo** contains samples with related events that provide alternative details for constructing event-grounded distractors. Importantly, both subsets probe long-range recall and are instantiated as single-turn QA, while EgoProbe-Echo further challenges models with event-specific recall and coarse-to-fine multi-turn QA.

The overall construction pipeline summarized in Fig. 2 illustrates how our data engine transforms source QAs into probes of long-range memory and event-specific recall. Starting from the source QA and its video and timestamped descriptions, the engine progressively transforms each sample through four automatic stages. 1) It first performs *Related-event Discovery* to identify whether the video contains events related to the target event, which determines whether the sample follows the EgoProbe-Base or EgoProbe-Echo track. 2) It then performs *Recall Interval Extension* by delaying the query time to increase the temporal distance from the target event. 3) *Query Disambiguation* further checks and refines the question to ensure that it still uniquely refers to the original target event while preserving the original answer. 4) *Multiple-choice and Dialogue Generation* converts the refined QA into a four-option multiple-choice question; for EgoProbe-Echo samples, this stage additionally introduces event-grounded distractors and derives coarse-to-fine two-turn dialogues. Finally, *Human Verification* checks the generated QA and associated video evidence to ensure the accuracy and quality of the final benchmark. Note that all resulting probes are defined under a streaming protocol: models process video observations chronologically and maintain memory for subsequent recall, and the query is revealed only at its designated arrival time t_q.

Step 1: Related Event Discovery. To identify recurring event patterns, we first use lexical retrieval to search the timestamped video descriptions after the target event $e^\star$ for other events sharing question-relevant objects, actions, or activities. An LLM-based module then verifies whether each candidate is a distinct related event and identifies its associated answer. If this answer differs from the target answer, the candidate is retained as a valid contrastive event, and its answer serves as an event-grounded distractor. We retain event timestamps to guide recall-interval extension in Step 2. In Step 4, alternative answers provide event-grounded distractors, while shared event patterns guide coarse-grained question generation.

Step 2: Recall Interval Extension. We extend the recall interval Δt by delaying query arrival according to the video's duration and event structure. For samples with adjustable query times, we select t_q from four predefined recall-interval ranges: short (< 3 min), medium (3–30 min), long (30 min–12 h), and ultra-long (≥ 12 h), subject to video boundaries. The timestamps of related events identified in Step 1 further guide query placement, determining which events fall within the resulting recall interval. This allows us to control both the temporal distance from the target event and the related events encountered by the model before answering.

Step 3: Query Disambiguation. Extending the recall interval may introduce ambiguity, as multiple events in the observed history could match the original question. To resolve this, we employ a feedback-guided disambiguation loop with three LLM-based roles: a *detector*, an *editor*, and a *verifier*. The detector first checks whether the question uniquely identifies the target event among all events before t_q. If ambiguity is detected, the editor refines the question by adding minimal cues grounded in the target-event evidence while preserving the original answer. The verifier then checks target-event identification, cue grounding, and answer leakage. Questions passing verification are retained; otherwise, the verifier's feedback guides refinement, and unresolved cases are discarded.

Step 4: Multiple-choice and Dialogue Generation. Given the refined question from Step 3 and its original answer, we convert the source open-ended QA into a four-option multiple-choice question to enable more consistent evaluation and better leverage the event-grounded distractors identified in Step 1. For EgoProbe-Base, an LLM generates three plausible incorrect options. For EgoProbe-Echo, we prioritize event-grounded distractors as hard negatives and use an LLM to generate additional incorrect options when needed. Each question contains one ground-truth answer and three wrong options, yielding single-turn VQA instances for both tracks.

For EgoProbe-Echo, we additionally derive coarse-to-fine two-turn dialogues to probe different levels of memory recall. We use the semantic pattern shared by the target and related events to formulate the first turn R1 as a yes/no question with an affirmative ground-truth answer. The second turn R2 adapts the corresponding single-turn question into a follow-up that retains its target-event cues, correct answer, and four options, requiring the model to retrieve the detail associated with the designated event rather than a competing detail from another observed event. A separate verification call checks shared-event support, answer leakage, and preservation of the original detail question before accepting the dialogue. Further construction details are provided in Appendix B.

Human Verification. Trained human annotators review every automatically constructed QA and its associated video evidence through a dedicated web interface. They verify question clarity, target-event identification, answer support and uniqueness, event–distractor associations, and consistency between single-turn questions and derived dialogues. More details are provided in Appendix C.

3.3 Statistics and Coverage

EgoProbe-Base contains 1,137 single-turn QA pairs, and EgoProbe-Echo contains 1,229 single-turn QA pairs with 1,229 corresponding coarse-to-fine two-turn dialogues, yielding 2,366 independent QA pairs across 671 egocentric videos. Figure 3 summarizes the benchmark's temporal coverage, queried information types, related-event distribution, and distractor composition. We divide recall intervals into four categories: Short ($<$ 3 min), Medium (3–30 min), Long (30 min–12 h), and Ultra-long ($\geq$ 12 h). The intervals range from 31.63 seconds to 138.54 hours, with 24.4% of QA pairs requiring recall after at least 30 minutes. Questions cover location, entity identity, object source and placement, activity and event context, and attributes and states. Within EgoProbe-Echo, 58.6% of QA pairs have multiple annotated related events, and 30.4% contain two or three event-grounded distractors. These distributions cover different recall intervals and related-event configurations, sup-

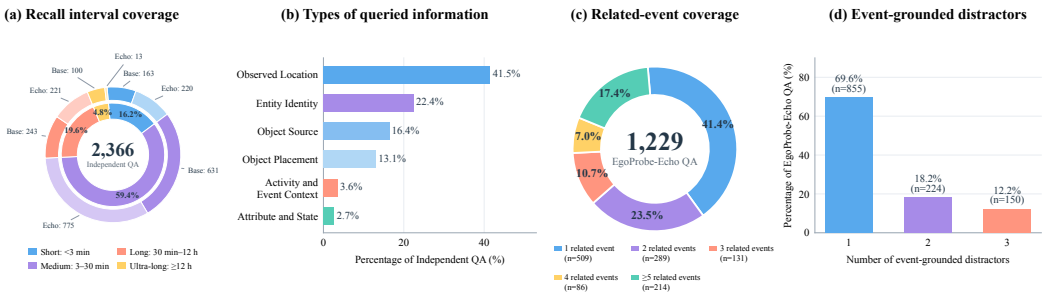


Figure 3: **EgoProbe statistics**: (a) recall intervals and (b) queried information types across EgoProbe; (c) related events and (d) event-grounded distractors per question across EgoProbe-Echo.

Table 2: Overall accuracy (%) on EgoProbe. Bold indicates best within each model category. Row colors indicate offline evaluation ( ) and streaming approaches based on recent windows ( ), KV-cache management ( ), event-centric memory ( ), or streaming reasoning and interaction ( ).

Model	Backbone	EgoProbe-Base	EgoProbe-Echo		
		Single-turn	Multi-turn		Single-turn
		Acc.↑	R1 Acc.↑	R2 Acc.↑	Acc.↑
Proprietary MLLMs					
Gemini-3.5-Flash	–	51.10	79.25	33.77	**43.69**
GPT-5.6-Sol	–	**55.32**	**89.75**	**43.69**	43.37
Open-source General-purpose MLLMs					
Qwen3.5	Qwen3.5-9B	45.91	81.53	34.83	35.31
InternVL3.5	Qwen3-8B	47.58	86.66	30.43	34.91
LLaVA-OneVision-2	Qwen3-8B	**52.07**	**93.82**	**38.24**	**42.96**
Open-source Long-video Models					
Ego-R1	Qwen2.5-3B-Instruct	40.99	43.53	13.10	32.14
EgoRAG	Qwen3-32B	36.50	44.18	15.62	31.57
VideoChat-Flash	Qwen2-7B	**53.21**	92.68	38.65	41.66
EM^{2}Mem	Qwen3.5-35B-A3B	47.93	**93.65**	**42.72**	**44.10**
Streaming Models					
SimpleStream	Qwen3-VL-8B	28.41	25.71	6.75	27.45
StreamingVLM	Qwen2.5-VL-7B	44.50	41.09	12.69	31.24
HERMES	Qwen2.5-VL-7B	46.44	61.19	21.56	34.17
StreamForest	Qwen2-7B	**50.66**	88.85	**35.39**	**38.89**
OASIS	Qwen3-VL-8B	47.41	70.55	25.06	36.21
EventMemAgent	Qwen3-VL-8B	49.52	88.36	32.63	37.75
VST	Qwen2.5-VL-7B	48.64	**95.28**	33.93	35.48
JoyAI-VL-Interaction	JoyAI-VL-1.0 (8B)	45.29	81.86	26.53	31.49

porting evaluation of long-range memory and event-specific recall. Representative QA examples from both subsets are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Evaluated Models. We evaluate 17 models across four categories: (1) *proprietary MLLMs*, including Gemini-3.5-Flash Google DeepMind (2026) and GPT-5.6-Sol OpenAI (2026); (2) *open-source general-purpose MLLMs*, including Qwen3.5 Team et al. (2025), InternVL3.5 Wang et al. (2025), and LLaVA-OneVision-2 Li et al. (2024a); (3) *long-video models*, including Ego-R1 Tian et al. (2025), EgoRAG Yang et al. (2025), VideoChat-Flash Li et al. (2026a), and EM^{2}Mem Chen et al. (2026); and (4) *streaming models*, covering representative approaches with different design emphases: recent-frame windows (SimpleStream Shen et al. (2026)), KV-cache management (StreamingVLM Xu et al. (2026) and HERMES Zhang et al. (2026)), event-centric memory (StreamForest Zeng et al. (2026), OASIS Liang et al. (2026), and EventMemAgent Wen et al. (2026)), and streaming reasoning and interaction (VST Guan et al. (2026) and JoyAI-VL-Interaction Yao et al. (2026)). Model-specific configurations are detailed in Appendix D.

Evaluation Protocols. All models are restricted to video observations available up to query time t_q. Under the *strict streaming* protocol, models process observations continuously without knowing the query and answer using their retained memory state. Long-video models instead adopt *offline prefix processing*, applying their native encoding or retrieval procedures to the available video history, while general-purpose and proprietary MLLMs receive frames sampled from this history within

their input budgets. For two-turn dialogues, both rounds share the same video evidence, with no additional observations introduced before R2.

Evaluation Metrics. Accuracy (%) is used as primary metric. Single-turn accuracy is the proportion of correctly answered questions. For two-turn dialogues, critically, R2 is counted as correct only when both rounds are answered correctly.

4.2 Overall Performance

Table 2 compares model families on long-range recall in EgoProbe-Base and event-specific recall in EgoProbe-Echo under their respective evaluation protocols.

Overall difficulty. EgoProbe remains challenging across model families. Uniform random guessing yields 25% single-turn accuracy and 12.5% two-turn completion accuracy when guessing independently across both rounds. Even the best model, GPT-5.6-Sol, achieves only 55.32% Base accuracy and 43.69% two-turn completion accuracy. Notably, SimpleStream achieves only 6.75% two-turn accuracy despite its strong results on OVO-Bench and StreamingBench, highlighting the limitations of recent-frame input and supporting the need for long-range memory.

Comparison across model families. Proprietary MLLMs perform strongly overall, while open-source general-purpose models remain competitive: LLaVA-OneVision-2 reaches 52.07% on EgoProbe-Base. Among long-video models, VideoChat-Flash achieves 53.21% Base accuracy, and EM^{2}Mem reaches 42.72% two-turn accuracy. These results use query-time access to historical video, unlike memory maintained during streaming observation. Among streaming models, StreamForest leads on both tasks with 50.66% and 35.39%, respectively. The remaining gap highlights the difficulty of supporting reliable recall under streaming constraints.

Recall among related experiences. All models obtain lower single-turn accuracy on EgoProbe-Echo than on EgoProbe-Base. Although the subsets are not matched controls, this consistent pattern indicates greater difficulty in answering about target-event details among related experiences.

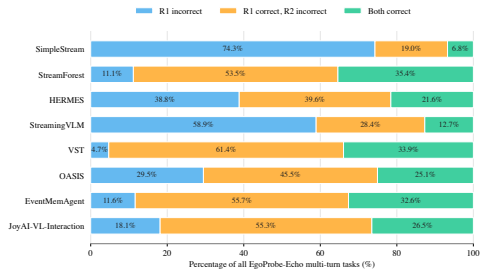
Coarse versus precise recall. The coarse-to-fine results further expose this challenge: VST achieves 95.28% R1 accuracy but only 33.93% two-turn completion accuracy, showing that a correct coarse response does not ensure accurate target-detail recall.

Single-turn versus two-turn performance. Echo single-turn accuracy exceeds two-turn completion accuracy for most models, indicating that success on a standalone detail question does not necessarily carry over to the two-turn task. This comparison motivates examining models' ability to answer consistently across turns, while also accounting for failures at the initial R1 stage.

Table 3: EgoProbe-Base accuracy (%) across recall intervals. Bold marks column maxima.

Model	Short < 3 min	Medium 3–30 min	Long 30 min–12 h	Ultra-long ≥ 12 h
StreamForest	50.92	**53.72**	**48.15**	37.00
HERMES	46.01	50.40	44.03	28.00
StreamingVLM	45.40	45.17	47.74	31.00
VST	**53.99**	50.55	44.03	**39.00**
OASIS	52.15	50.55	42.39	32.00
EventMemAgent	46.63	52.93	**48.15**	36.00
JoyAI-VL-Interaction	48.47	48.97	38.68	33.00

Figure 4: Two-turn outcomes on EgoProbe-Echo (% of all dialogues).



4.3 Long-Range Recall

We examine recall intervals on EgoProbe-Base, whose verified intervals contain no intervening related events. This separates recall interval from related-event interference, allowing us to examine how performance varies with the time elapsed since the target event. Table 3 reports streaming-model accuracy across the four recall-interval groups defined in Section 3.3. Every model attains its lowest accuracy in the Ultra-long group, where scores range from 28.00% to 39.00%. These re-

Table 4: Single-turn answer outcomes (%) on EgoProbe-Echo. Event err. and Other err. denote event-linked and other-option error rates, respectively. Bold marks the highest accuracy and lowest error rates among models within each event group. The gray row denotes random choice.

	1 event			2 events			≥3 events		
Model	**Acc.↑**	**Event err.↓**	**Other err.↓**	**Acc.↑**	**Event err.↓**	**Other err.↓**	**Acc.↑**	**Event err.↓**	**Other err.↓**
Random choice	25.00	25.00	50.00	25.00	36.93	38.07	25.00	53.01	21.99
StreamingVLM	37.08	35.47	27.45	22.41	52.28	25.31	27.12	63.56	9.32
HERMES	36.44	36.44	27.13	31.54	46.89	21.58	**32.05**	**60.55**	7.40
StreamForest	43.18	38.04	**18.78**	**39.83**	**44.40**	15.77	30.96	61.64	7.40
OASIS	42.54	34.19	23.27	29.88	56.02	**14.11**	29.59	65.21	**5.21**
EventMemAgent	**44.46**	**31.94**	23.60	33.61	46.89	19.50	29.04	63.84	7.12
VST	42.38	37.72	19.90	30.29	53.11	16.60	27.12	66.85	6.03
JoyAI-VL-Interaction	36.28	36.76	26.97	29.05	51.87	19.09	24.93	65.21	9.86

sults highlight the difficulty of recalling distant target events even without intervening related events. Figure 9 in Appendix E.1 visualizes accuracy across recall intervals.

4.4 Event-Specific Recall under Interference

We use EgoProbe-Echo single-turn QA to analyze whether models recall the target-event answer, select a detail from another related experience, or select another incorrect option. Table 4 groups questions by one, two, or at least three related events supporting alternative answers. Alongside accuracy, we report the *event-linked error rate* and *other-option error rate*, defined as the proportions of all questions answered with event-grounded and other incorrect options, respectively. The random-choice row accounts for the distribution of these option types within each group.

Related events challenge precise recall and reveal event-linked errors. All models achieve lower accuracy in multiple-event groups than in the single-event group. For example, StreamForest's accuracy decreases from 43.18% with one related event to 39.83% with two and 30.96% with at least three. Meanwhile, event-linked error rates increase across the three groups for every model and consistently exceed their random-choice references. These results highlight difficulty of recalling the correct detail when similar experiences contain different answers. EgoProbe captures this challenge by identifying answers that are valid for another event but incorrect for the queried event.

Coarse-to-fine probing distinguishes failure levels. While event links characterize the details selected in incorrect answers, two-turn probing examines the level of recall at which responses fail. Figure 4 decomposes all EgoProbe-Echo dialogues into three outcomes: incorrect R1, correct R1 followed by incorrect R2, and both rounds correct. For VST, 61.4% of all dialogues have a correct R1 response but an incorrect R2 response. This outcome also accounts for over half of dialogues for StreamForest, EventMemAgent, JoyAI-VL. Staged probing therefore distinguishes failure to answer the coarse question from failure to recall the target-specific detail after a successful coarse response.

5 Conclusion and Limitations

We introduce EgoProbe, a benchmark for long-range memory and event-specific recall in streaming egocentric video. Through an event-grounded data engine, EgoProbe constructs memory probes that evaluate both how far models can recall and how exact they recover target-event details. Our evaluation reveals substantial limitations in both dimensions: models struggle to recall distant events and recover target-specific details. EgoProbe provides a testbed for developing and diagnosing memory capabilities in streaming egocentric systems.

Limitations. Our framework supports longer streams, but EgoProbe's temporal coverage is limited by the available EgoLife recordings. Future work will explore recordings spanning weeks or months to evaluate longer-term memory. EgoProbe also omits audio information, an important component of egocentric experiences; incorporating it would enable a more comprehensive evaluation of egocentric memory.

AI USE STATEMENT

We used LLMs in benchmark construction to assess related events, disambiguate questions, generate answer options, derive two-turn dialogues, and verify generated candidates. Professional annotators reviewed the resulting QA items and associated video evidence. We also used AI tools to assist with paper writing. The authors reviewed the AI-assisted content and take responsibility for the final paper, dataset, and reported findings.

ETHICS STATEMENT

EgoProbe is constructed from existing Ego4D and EgoLife recordings, without collecting new participant videos. Egocentric recordings may contain identifiable individuals and private environments; any use or redistribution of the source videos must follow the respective datasets' licenses and access conditions. Professional annotators reviewed the generated QA and associated evidence using a dedicated interface and explicit guidelines to assess question clarity, answer validity, and event associations. EgoProbe is intended to study memory in egocentric video understanding. Potential misuse includes unauthorized recall of personal activities, and benchmark performance should not be interpreted as establishing suitability for deployment in privacy-sensitive settings.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the benchmark construction pipeline, with implementation details and prompts provided in Appendix B and human-review guidelines in Appendix C. The experimental setup specifies model families, history-access protocols, and evaluation metrics, with additional model configurations provided in Appendix D.2. These details document how the QA instances are constructed and how model responses are evaluated, including the correspondence between single-turn questions and their derived two-turn dialogues.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Leonard Bärmann and Alex Waibel. Where did i leave my keys?—episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1559–1567. IEEE, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Yijun Chen, Yaqi Zheng, Yanya Li, Boyi Xiao, Buqiang Xu, Shuofei Qiao, Jizhan Fang, Xinle Deng, Yunzhi Yao, Xuehai Wang, et al. Emˆ 2mem: Event-centric multimodal memory for large language models. *arXiv preprint arXiv:2609.00551*, 2026.

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.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. Egostream: A diagnostic benchmark for streaming episodic memory in egocentric vision. *arXiv preprint arXiv:2605.31557*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Google DeepMind. Gemini 3.5 Flash Model Card, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf.

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.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Yiran Guan, Liang Yin, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Video streaming thinking: Videollms can watch and think simultaneously. In *European Conference on Computer Vision*, pp. 80–99. Springer, 2026.

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. IEEE, 2025.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. Egosat: A comprehensive benchmark of egocentric streaming interaction understanding. In *European Conference on Computer Vision*, 2026.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026a.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6592–6600, 2026b.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. *arXiv preprint arXiv:2604.17052*, 2026.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 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.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

OpenAI. GPT-5.6 Sol. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913. IEEE, 2025.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24129–24138. IEEE, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

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

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. River: A real-time interaction benchmark for video llms. *arXiv preprint arXiv:2603.03985*, 2026.

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

Qwen Team et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 6(7):13, 2025.

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.

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*, 2025.

Xiaodong Wang, Xuanyi Zhao, Pedro Rodriguez, Devendra Singh Sachan, Barlas Oguz, Seungwhan Moon, Shang-Wen Li, Gargi Ghosh, Xin Dong, and Wen-Tau Yih. S-ember: A large-scale benchmark for streaming egocentric memory retrieval. *arXiv preprint arXiv:2607.02689*, 2026.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use. *arXiv preprint arXiv:2602.15329*, 2026.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

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.

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

Haowei Zhang, Shudong Yang, Jinlan Fu, See Kiong Ng, and Xipeng Qiu. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8411–8430, 2026.

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 preprint arXiv:2406.16852*, 2024.

A BENCHMARK EXAMPLES

The following examples illustrate long-range recall in EgoProbe-Base and event-specific recall in EgoProbe-Echo. Each panel presents selected video frames, followed by the question and answer options. For Echo, the options are shared by the single-turn question and R2. Event-grounded distractors are supported by other related experiences but remain incorrect for the target event; other options have no such event association.

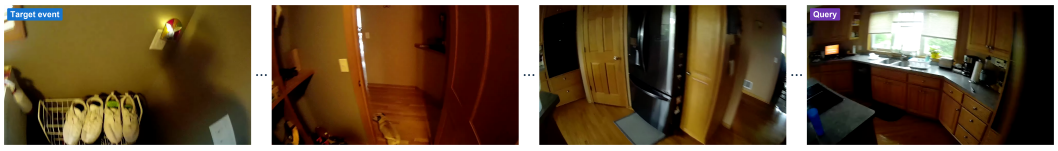


Figure 5: EgoProbe-Base single-turn example. The model must recall the earlier location of the shoes after approximately 17 minutes.

EgoProbe-Base: recalling the location of shoes. Target evidence: 12:15–12:17; query time: 29:19; recall interval: 17 min 2 s.

Single-turn question: In what location did I see my shoes?

Option	Answer	Role
A	inside the shoe cabinet	Other option
B	under the bed	Other option
C	on the floor beside the door	Other option
D	**on the shoe rack**	**Correct answer**

The target evidence shows the shoes on a shoe rack. The question is asked after subsequent observations, with no intervening related event in the verified recall interval. All three incorrect options are other options rather than event-grounded distractors.

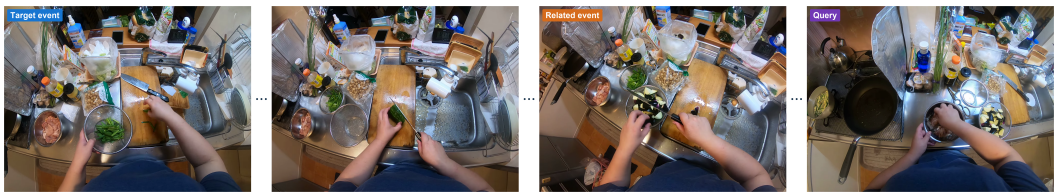


Figure 6: EgoProbe-Echo example with one event-grounded distractor. R1 asks about shared event content, while the single-turn question and R2 require the target-specific ingredient.

EgoProbe-Echo: recalling food placed in a sieve. Target evidence: 02:59.09–03:12.50; query time: 11:17; recall interval: 8 min 4.50 s.

Single-turn question: What did I put in the sieve right after picking it from the dish drainer?

Two-turn dialogue. R1: Did I put food into the sieve? **Answer: Yes** (choices: Yes / No).
R2, after Yes: What did I put in the sieve right after picking it from the dish drainer?

Option	Answer (single-turn and R2)	Role
A	sago	Other option
B	cabbage	Other option
C	**spring onion**	**Correct answer**
D	eggplant	Event-grounded distractor

The wearer picks up the sieve from the dish drainer at 02:59.35 and puts spring onion into it at 03:00.42. A later event at 07:57.58 involves putting eggplant into the sieve, supporting option D. Both the target and the eggplant event support Yes in R1; the cue "right after picking it from the dish drainer" identifies the target for R2.



Figure 7: EgoProbe-Echo example with two event-grounded distractors. The wooden-plank and pocket locations are supported by later related events but are incorrect for the event preceding use of the gauge.

EgoProbe-Echo: recalling where a green pencil was placed. Target evidence: 27:16.95–27:19.16; query time: 01:06:07; recall interval: 38 min 47.84 s.

Single-turn question: Where did I put the green pencil before using the gauge?

Two-turn dialogue. R1: Did I place a pencil? **Answer: Yes** (choices: Yes / No).
R2, after Yes: Where did I put the green one before using the gauge?

Option	Answer (single-turn and R2)	Role
A	in my pocket	Event-grounded distractor
B	**on the wall**	**Correct answer**
C	on the wooden plank	Event-grounded distractor
D	on the table	Other option

The target event places the pencil on the wall before the wearer uses the gauge at 27:20.00. Later, the pencil is placed on a wooden plank at 29:45.62 and in a pocket at 50:06.03, supporting options C and A, respectively. These experiences all support Yes in R1, but only the wall location answers the target-specific question. R2 uses "the green one" to refer back to the pencil in R1. For consistent first-person wording in this illustration, option A is displayed as "in my pocket" instead of the source wording "in his pocket."

B Detailed Benchmark Construction

This section expands the four stages in Section 3.2. We preserve source QA and target evidence as construction anchors, while recording query-time decisions, question revisions, and event–option associations. Human verification is described separately in Appendix C.

B.1 Source Data

For QaEgo4D, we use QA aligned with Ego4D temporal evidence and timestamped narrations. For EgoLife, we use episodic QA and dense video descriptions from the A1_JAKE stream. EgoLife questions marked as requiring audio are excluded from the visual-only construction pool. The English description index preserves the original caption identifiers and timestamps, allowing retrieved descriptions to be traced back to video evidence. We use *video descriptions* for both sources throughout the paper. Each construction record contains the source question, canonical answer, target evidence, and stream identity. Target evidence may comprise multiple segments; the latest evidence endpoint determines the recall interval.

B.2 Related-Event Discovery

We retrieve descriptions using question-relevant entities, actions, and relations, with lexical matching and compatible aliases. QaEgo4D searches after the target evidence within the source video; EgoLife searches between the target evidence and the fixed native query. Nearby descriptions are grouped into candidate events rather than counted independently. In our setting, the QaEgo4D configuration groups descriptions separated by at most 8 seconds into windows of at most 30 seconds.

An LLM then assesses whether each candidate describes a distinct related experience and whether its evidence supports an answer to the requested detail. Related events with the same answer are retained as context but do not supply alternative answer options. Overlapping descriptions of the same experience are consolidated, retaining their supporting evidence. Each retained event therefore has a temporal reference, a relationship to the target, and an answer association when supported. Event counts refer to these annotated, deduplicated experiences.

B.3 Recall-Interval Extension and Sample Selection

Query-time planning selects a legal time after the target and the included related-event evidence, bounded by the video endpoint and, where applicable, the next related event outside the selected context. Sampling balances temporal coverage and single- versus multiple-related-event conditions, with a per-batch cap of five targets from one video.

B.4 Question Disambiguation

For adjustable questions, a detector evaluates whether the original wording uniquely identifies the target among the events preceding the planned query. Clear questions bypass rewriting. For ambiguous questions, an editor introduces only necessary cues supported by the target's local evidence, preserving the requested fact and answer. A verifier checks the revised reference, cue support, and answer leakage. Failed checks are returned to the editor, with at most two rewrite attempts; unresolved candidates are excluded.

Two retained examples illustrate how evidence-supported cues resolve event ambiguity. Both have complete detection, rewriting, and verification records.

Example 1: Disambiguation through an action cue

Original question	Where was the spoon before I picked it?
Refined question	Where was the spoon before I picked it **and cleaned it with a tissue**?
Ambiguity	The wearer picks up a spoon from a plate (47:29.97), a baking table (48:45.20), and a tray (52:04.10).
Grounded cue	Cleaning the spoon with a tissue (47:31.94) identifies the target occurrence without revealing its prior location.
Correct answer	A. in a plate (unchanged)
Event-grounded distractors	C. on a tray; D. on the baking table
Other option	B. in the cabinet

Example 2: Disambiguation through a location cue

Original question	What hand tool did I pick?
Refined question	What hand tool did I pick **from the wall**?
Ambiguity	The wearer picks up an air pump from a wall (18:09.16), a wrench from a machine (20:23.66), and a putty knife from a lift stand (21:01.87).
Grounded cue	The wall location identifies the target occurrence while leaving the tool's identity to be recalled.
Correct answer	D. air pump (unchanged)
Event-grounded distractors	B. wrench; C. putty knife
Other option	A. pliers

B.5 Multiple-Choice and Dialogue Generation

Every accepted single-turn task has one correct answer and three incorrect options. For EgoProbe-Base, the construction module generates or selects plausible alternatives of the required answer type without attaching intervening related-event evidence. For EgoProbe-Echo, answers supported by other related events are prioritized as event-grounded distractors, and remaining slots are filled with other options. EgoLife can reuse its original incorrect choices. Option verification checks distinctness, semantic equivalence, answer support, and whether an alternative is still incorrect for the specified target.

Two-turn generation operates on accepted EgoProbe-Echo questions. A shared semantic anchor is extracted from the target and selected related events, and R1 asks whether the corresponding experience occurred. R2 is a conversational adaptation of the single-turn question with its target cues, correct answer, and options preserved. A separate verification call checks support across the

selected events, absence of answer leakage, and preservation of the detail task. Parent and dialogue records retain their correspondence and share the same query time and visual evidence.

B.6 Models, Prompts, and Implementation Settings

The construction stage employs `gemini-3.7-flash` for all LLM roles. QaEgo4D semantic assessment and dialogue construction use the medium thinking setting; EgoLife single-turn processing uses the low setting and dialogue construction uses medium. The generation limit is 16,384 output tokens for these stages. Structured responses are checked against the expected fields, and verification is performed in a separate call.

Table 5 provides a concise overview of each role's core responsibilities and key prompt instructions, rather than the full prompts. Each call additionally supplies the relevant question and evidence records with stable event identifiers; dynamic sample payloads and output schemas are specified separately from the role instructions.

Table 5: Core instructions used by the construction roles.

Role	**Instruction and constraint**
Event assessment	Determine whether the candidate is a distinct related event. Extract an answer only when its own descriptions support the queried detail; retrieval similarity is not evidence.
Clarity detector	Apply the full question to the supplied pre-query events. Identify alternative plausible referents without answering or rewriting the question.
Question editor	Add minimal target-grounded cues that distinguish the target. Preserve the queried fact and correct answer, and do not reveal that answer.
Revision verifier	Check target identification, evidence for added cues, and answer leakage. Return specific feedback when a revision fails.
Option construction	Prioritize safely supported related-event answers, fill the remaining slots, and reject equivalent options or additional correct answers.
Dialogue construction	Identify shared content supported by the selected events, ask a coarse yes/no question, and adapt the original detail question into a follow-up.
Dialogue verification	Check shared-event support, coherence, answer leakage, and preservation of the original answer and options.

C Human Verification

C.1 Review Interface and Evidence

Professional annotators use a dedicated web interface to inspect the target, related-event, and query clips together with their source timestamps (Figure 8). The interface displays the question, options, original answer label, and event–option links, and supports separate single-turn and multi-turn decisions. Annotators are instructed to assess all supplied evidence rather than assume that the original answer or target/related-event labels are correct.

C.2 Annotation Guidelines

Review follows six criteria: (1) the question clearly identifies the intended event, entity, and temporal relation; (2) the target video supports the labeled answer; (3) the options are distinct and exactly one answers the question; (4) the derived dialogue preserves the single-turn meaning and remains coherent across rounds; (5) EgoLife person references are understandable from visual evidence, with names replaced by unambiguous descriptions when appropriate; and (6) each event-linked incorrect option is supported by its associated clip while remaining incorrect for the target. Other options do not require supporting related-event clips.

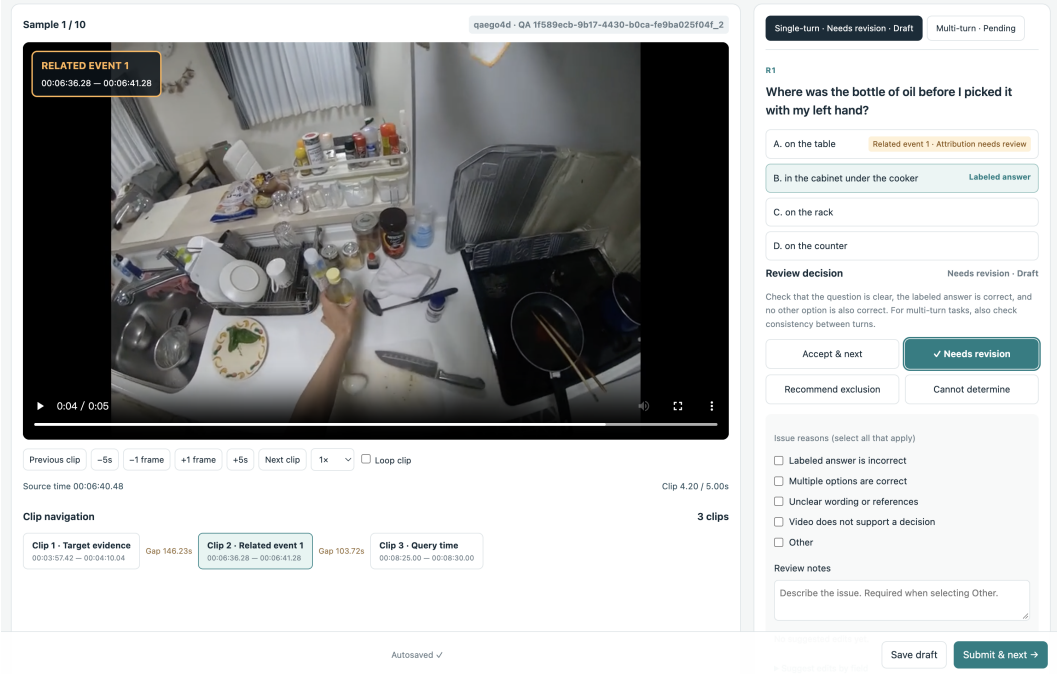


Figure 8: Human-review interface. Annotators navigate timestamped evidence clips, check QA and event-linked options, and record decisions with reasons and suggested edits. Labels shown in the interface are subject to verification.

C.3 Review and Adjudication Procedure

Annotators select *Accepted*, *Needs revision*, *Recommend exclusion*, or *Cannot determine*, and provide reasons and field-level suggestions when needed. A single-turn question and its derived dialogue form one review unit, but their decisions are submitted separately; both dialogue rounds are reviewed together. Suggested edits are kept separate from the original QA until they are applied and checked.

The interface-based review checks QA quality and event correspondence in the supplied clips. Subsequent human verification reviews the full video to confirm the absence of related events within the EgoProbe-Base recall intervals. Claims that require broader temporal context, such as first or last occurrences, are checked against the relevant history before acceptance.

C.4 Review Coverage and Outcomes

Before human filtering, the candidate pool contains 1,264 EgoProbe-Base single-turn questions and 1,291 EgoProbe-Echo single-turn questions, each of the latter paired with a two-turn dialogue. This gives 2,555 independent QA instances and 3,846 evaluation tasks.

After human review and filtering, the final benchmark contains 1,137 Base questions, 1,229 Echo questions, and 1,229 corresponding dialogues: 2,366 independent QA instances and 3,595 evaluation tasks. The retained task count is 93.5% of the pre-filtering pool, with a net reduction of 251 tasks. This is the final retention rate, including items accepted after revision. Annotators record issues and suggested corrections, and revised items are rechecked before final acceptance.

D Evaluation Protocols and Implementation Details

D.1 History-Access Protocols

All settings enforce the query-time cutoff: model inputs exclude observations after t_q, and benchmark target timestamps, answer labels, construction descriptions, and event–option links are withheld from the model. History access nevertheless differs across implementations. To enforce strict streaming evaluation, streaming models process observations chronologically without access to future questions and answer each query using only their retained memory and explicitly stored observations, without reloading discarded video. Prefix-processing references sample or retrieve permitted historical information at query time.

We adapt models where needed to enforce strict streaming evaluation while preserving their original memory mechanisms. For VST, frames are sampled and grouped into fixed-duration chunks independently of query arrival. Each completed chunk updates textual memory through the official thought-generation procedure, after which old frames are released except for a recent-video cache. Query arrival does not trigger an additional persistent memory update. At query time, the model answers using this memory and the available recent frames in an isolated branch. R1 and R2 share dialogue context, but their responses are not written back to the persistent memory, allowing observation to continue independently of the questions. For EventMemAgent, we use incremental event-memory construction: new observations update the retained event memory throughout the stream, and queries are answered from the accumulated state without rebuilding memory from the full video prefix.

D.2 Model Configurations

Table 6 summarizes the models run configurations. Model-specific processing budgets are retained rather than equated across architectures. General-purpose visual baselines use 128 sampled prefix frames; VideoChat-Flash uses a larger budget of up to 512 frames. API requests use 128 frames resized to a maximum side of 448 pixels. We evaluate Gemini-3.5-Flash and GPT-5.6-Sol; the API answer-generation limit is 8,192 tokens.

Table 6: Input and memory settings recovered from evaluation records. Frame budgets describe the available input, not the full duration of the source stream.

Model	**Input and principal settings**
Gemini / GPT	128 prefix frames; maximum side 448 pixels; high image detail.
Qwen3.5	128 prefix frames; per-frame pixel budget 200,704; non-thinking.
InternVL3.5	128 prefix frames; one 448×448 patch per frame.
LLaVA-OneVision-2	128 prefix frames; frame input rather than codec input.
Ego-R1	1 FPS; 30-frame caption chunks; agent limit of seven turns; temporal video retrieval.
EgoRAG	1 FPS; 30-frame caption chunks; Qwen3-32B text retrieval/answering.
VideoChat-Flash	Native dynamic sampling, up to 512 prefix frames; 448-pixel model.
EM^{2}Mem	Precomputed memory graphs; retrieval limits: 5 episodic, 8 semantic, and 3 visual items; up to three rounds.
SimpleStream	1 FPS; a recent window of four frames in the available run configuration.
StreamingVLM	2 FPS; two-frame chunks; visual window 16; text sink and sliding window 512 each.
HERMES	1 FPS; 16-frame chunks; KV budget 6,000.
StreamForest	1 FPS; native online generation and memory update.
OASIS	2 FPS; buffer limit 32 frames; current window 16; up to 16 frames per event node; event retrieval limit 2.
EventMemAgent	Incremental event-memory construction at 1 FPS; up to ten assistant turns; Qwen3-VL-4B caption service.
VST	Query-independent fixed chunks; official thought-based textual memory; isolated answers using memory and recent frames.
JoyAI-VL-Interaction	Journals record chronological observation, eight-frame input chunks, and retained hierarchical summaries.

Qwen3.5 uses non-thinking generation with temperature 0.7, top-p 0.8, top-k 20, and a 16,384-token output limit. LLaVA-OneVision-2 uses deterministic decoding with a 1,024-token limit. HERMES uses temperature zero and a 16-token answer limit; StreamingVLM uses option-logit selection for the final answer.

D.3 Evaluation Procedure

Predictions are joined to the fixed evaluation lists by sample identifier: 1,137 EgoProbe-Base single-turn questions, 1,229 EgoProbe-Echo single-turn questions, and their 1,229 derived dialogues. The source-question and dialogue correspondence is preserved throughout evaluation.

In the shared streaming harness, queries are processed in chronological order. Each QA is answered in an isolated branch of the current state, preventing one question or its answer from altering the persistent memory used by another. An incomplete observation chunk at the query boundary is handled within that branch without consuming future frames. For a dialogue, the actual R1 response is retained as conversational context for R2. A parsed affirmative response triggers R2; a negative or invalid R1 response does not. Both rounds use the same visual evidence, and only their dialogue context changes.

D.4 Metrics and Scoring

Single-turn accuracy is the number of correct answers divided by the fixed number of questions. R1 accuracy uses all dialogues as its denominator. R2 accuracy also uses all dialogues and counts a task as correct only when both rounds succeed; dialogues that never reach R2 contribute zero. The released R1 questions are constructed from observed events and have affirmative labels. The 12.5% two-turn chance reference therefore denotes independent uniform guessing over the two R1 choices and four R2 choices, not an always-affirmative baseline.

For event-linked analysis, let N be the number of questions in a group and k_i the number of distinct event-grounded incorrect options for question i. Event-linked error rate counts selections of those options and divides by N. Its uniform-choice reference is $100\sum_i(k_i/4)/N$. Other-option errors count selections of the remaining incorrect options.

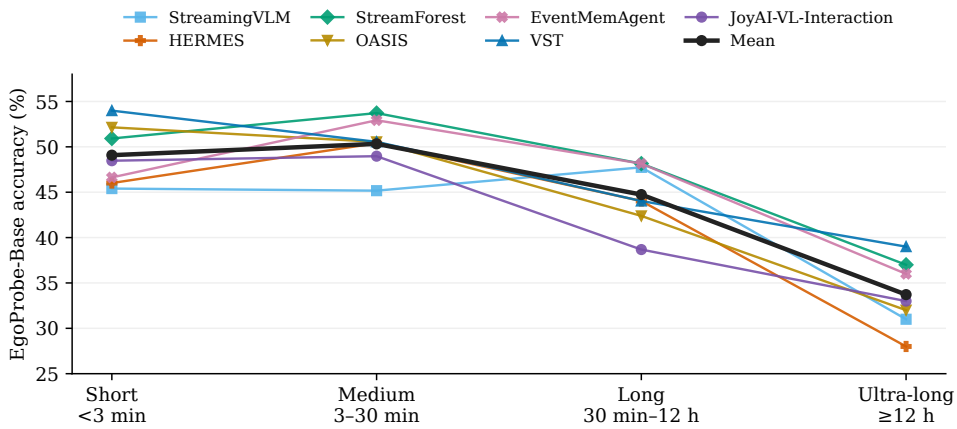


Figure 9: EgoProbe-Base accuracy across recall intervals, using the seven models in Table 3. The black curve is their unweighted mean. Connecting lines guide comparison between interval groups.

E Additional Experimental Results

E.1 Recall-Interval Analysis

Figure 9 visualizes the seven model entries in Table 3. The unweighted mean is 49.08% in Short and 50.33% in Medium, falling to 44.74% in Long and 33.71% in Ultra-long. All seven models

obtain their lowest accuracy in the Ultra-long group. Overall, accuracy tends to decrease as the recall interval grows, showing a pattern resembling a forgetting curve and highlighting the difficulty of retaining information for long-range recall.

E.2 TEXT-ONLY REFERENCE

Table 7 compares Qwen3.5 with and without visual input on the same EgoProbe-Base and EgoProbe-Echo single-turn questions. Given only the questions and answer options, the model achieves 25.51% and 24.49% accuracy, respectively, both close to the 25% random-choice baseline. This suggests that linguistic plausibility and prior knowledge alone offer limited help in identifying the correct answers for this model. Adding video raises accuracy to 45.91% on EgoProbe-Base and 35.31% on EgoProbe-Echo, corresponding to gains of 20.40 and 10.82 percentage points. The improvement on both subsets supports the importance of visual evidence, while the remaining errors show that access to this evidence does not ensure reliable recall. Together, these results support EgoProbe's evaluation of memory for observed experiences, where models must recover information from the video rather than rely on question and option text alone.

Table 7: Qwen3.5 single-turn accuracy (%) with and without video on EgoProbe-Base and EgoProbe-Echo single turn QAs.

Model / Input	**Base Acc.↑**	**Echo Acc.↑**
Qwen3.5 — Text-only	25.51	24.49
Qwen3.5 — Video + Text	45.91	35.31

E.3 FINE-GRAINED RELATED-EVENT ANALYSIS

Table 8 expands the event grouping in Table 4 into 1, 2, 3, 4, and at least 5 related events supporting different answers. The groups contain 623, 241, 125, 67, and 173 questions, respectively. We report single-turn accuracy and event-linked error rate with fixed group denominators.

Every model with reported group-level results has lower accuracy and a higher event-linked error rate in the at-least-five-event group than in the one-event group. For example, StreamForest changes from 43.18% accuracy and 38.04% event-linked errors to 31.79% and 63.58%, respectively. These results highlight the value of EgoProbe's event-grounded design: the benchmark evaluates target-specific recall across different numbers of related experiences and distinguishes errors linked to those experiences from other incorrect choices. This provides a more detailed view of recall failures than overall accuracy alone.

Table 8: Fine-grained EgoProbe-Echo single-turn results (%). Groups count annotated related events supporting different answers. Event err. denotes event-linked error rate. Bold indicates the highest accuracy and lowest error rate among reported models in each group.

	1 event		**2 events**		**3 events**		**4 events**		**≥5 events**	
Model	**Acc.↑**	**Event err.↓**	**Acc.↑**	**Event err.↓**	**Acc.↑**	**Event err.↓**	**Acc.↑**	**Event err.↓**	**Acc.↑**	**Event err.↓**
StreamingVLM	37.08	35.47	22.41	52.28	29.60	60.00	26.87	**58.21**	25.43	68.21
HERMES	36.44	36.44	31.54	46.89	37.60	**52.80**	25.37	64.18	30.64	64.74
StreamForest	43.18	38.04	**39.83**	**44.40**	31.20	60.00	**28.36**	59.70	31.79	63.58
OASIS	42.54	34.19	29.88	56.02	**40.00**	54.40	16.42	76.12	27.17	68.79
EventMemAgent	**44.46**	**31.94**	33.61	46.89	28.00	62.40	20.90	68.66	**32.95**	**63.01**
VST	42.38	37.72	30.29	53.11	27.20	64.80	14.93	77.61	31.79	64.16
JoyAI-VL-Interaction	36.28	36.76	29.05	51.87	26.40	62.40	23.88	61.19	24.28	68.79