EGOMEMREASON: A Memory-Driven Reasoning Benchmark for Long-Horizon Egocentric Video Understanding

Ziyang Wang[1,*] **Yue Zhang**[1,*] **Shoubin Yu**[1] **Ce Zhang**[1] **Zengqi Zhao**[1]
Jaehong Yoon[2] **Hyunji Lee**[1] **Gedas Bertasius**[1] **Mohit Bansal**[1]

[1]UNC Chapel Hill [2]NTU Singapore

Code Project page

Abstract

Next-generation visual assistants, such as smart glasses, embodied agents, and always-on life-logging systems, must reason over an entire day or more of continuous visual experience. In ultra-long videos, relevant information is sparsely distributed across hours or days, making **memory a fundamental challenge**: models must accumulate information over time, recall prior states, track temporal order, and abstract recurring patterns. However, existing week-long video benchmarks are primarily designed for perception and recognition, such as moment localization or global summarization, rather than reasoning that requires integrating evidence across multiple days. To address this gap, we introduce **EGOMEMREASON**, a comprehensive benchmark for week-long egocentric video understanding through memory-driven reasoning. EGOMEMREASON evaluates three complementary memory types: **entity memory**, tracking how object states evolve and change across days; **event memory**, recalling and ordering activities separated by hours or days; and **behavior memory**, abstracting recurring patterns from sparse, repeated observations over the whole week period. EGOMEMREASON comprises 500 questions across three memory types and six core challenges, with an average of 5.1 video segments of evidence per question and 25.9 hours of memory backtracking. We evaluate EGOMEMREASON on **17** methods across MLLMs and agentic frameworks, revealing that even the best model achieves only 39.6% overall accuracy. Further analysis shows that the three memory types fail for distinct reasons and that performance degrades as evidence spans longer temporal horizons, revealing that long-horizon memory remains far from solved. We believe EGOMEMREASON establishes a strong foundation for evaluating and advancing long-context, memory-aware multimodal systems.

1 Introduction

Next-generation visual assistants, from smart glasses (Grauman et al., 2022; 2024; Yu et al., 2026) to embodied agents (Hu et al., 2025b; Yang et al., 2025c; Zhang et al., 2024c) and always-on life-logging systems (Xu et al., 2025), must reason over continuous visual streams spanning an entire day or more. This has driven growing interest in long-form video understanding (Wang et al., 2025a; Nagrani et al., 2025; Yang et al., 2025b; Chandrasegaran et al., 2024), and more recently, in week-long video understanding (Yang et al., 2025a; Chen et al., 2026; Kim et al., 2026b; Yan et al., 2025). At this temporal scale, relevant information is sparsely distributed across hours or days, posing unique challenges that go well beyond those of short-clip or hour-long video understanding.

However, as video duration scales to days, densely sampling visual inputs (e.g., at 1 FPS) becomes impractical due to context length limitations, while the abundance of irrelevant content can further overwhelm models and hinder reasoning (Liu et al.,

*Equal contribution.

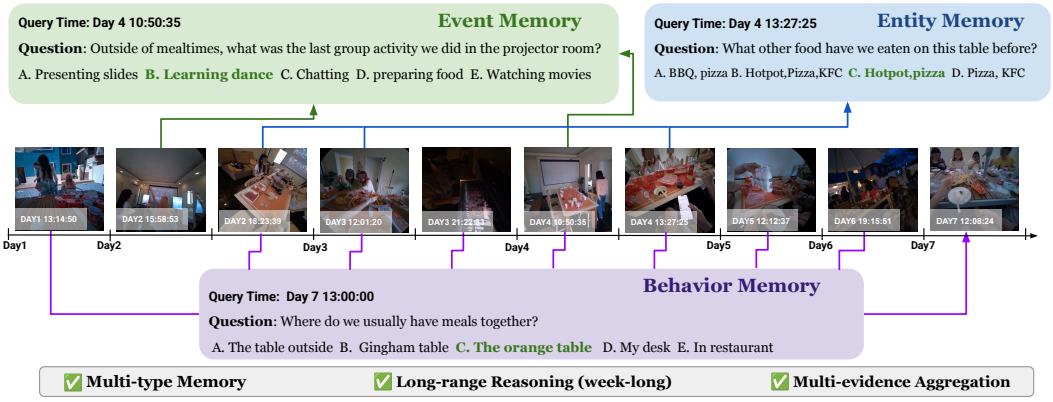

Figure 1: **Illustration of our EGOMEMREASON for week-long egocentric video memory.** Given a query at a specific time, answering requires retrieving and aggregating evidence from multiple temporally distant observations across days. We categorize memory into three types: entity memory (tracking persistent objects and states, tracking the same object for a long time), event memory (ordering and linking events, e.g. linking previous similar event details), and behavior memory (inferring patterns, e.g. activity habit).

2024). This makes **memory a fundamental challenge** for week-long video understanding: models must selectively accumulate information over time, recall previously observed states, track temporal order, and abstract recurring patterns from past experience. Yet existing benchmarks (Yang et al., 2025a; Tian et al., 2025; Yan et al., 2025) primarily target perception and recognition; their questions are typically answerable from a single moment (e.g., "What type of image did I paste into the slideshow?" (Yan et al., 2025)) or within a short temporal window of under ten minutes (e.g., "What was in the pot just before it was set aside?" (Yang et al., 2025a)), rather than requiring reasoning that accumulates multiple segments of evidence across hours to days. As shown in Figure 2, most existing benchmarks span limited temporal certification of backtracking to the ground-truth evidence (defined in EgoSchema (Mangalam et al., 2023)), and require only a small number of evidence segments per question. Taken together, these limitations suggest that current benchmarks do not yet capture the memory demands of week-long video understanding, and a benchmark designed for this setting remains an open challenge.

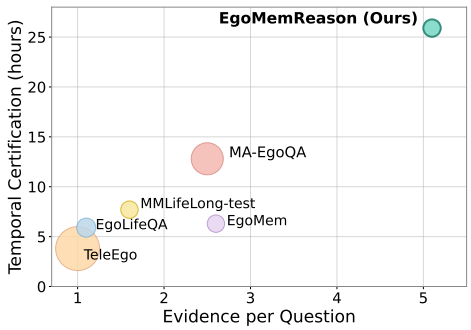

Figure 2: Comparison with existing week-long video benchmarks. The x-axis shows the average number of distinct video segments needed to answer a question (i.e., evidence), and the y-axis shows temporal certification in hours (i.e., the total video duration one must search to locate all ground-truth evidence). Bubble size is proportional to the number of questions.

To address this gap, we introduce **EGOMEMREASON**, a comprehensive benchmark designed to **systematically** evaluate week-long egocentric video understanding through the lens of memory-driven reasoning. While existing long-video benchmarks largely reduce question answering to locating one or a few relevant moments and performing localized reasoning, as shown in Figure 1, EGOMEMREASON requires models to aggregate, relate, and abstract over evidence distributed across days, operations that more closely resemble how humans actually reason over remembered experience. As shown in Figure 2, each question in EGOMEMREASON requires aggregating an average of 5.1 distinct video segments distributed over average 25.9 hours of memory backtracking, exceeding the strongest prior week-long benchmark by 2× in evidence count and 2× in temporal certification. To capture the breadth of memory required at this temporal scale, as shown in Figure 3, we decompose week-long memory into three complementary types, each targeting

a distinct reasoning operation over accumulated experience. *Entity* memory, which requires **aggregating** how objects evolve across days (e.g., recalling all foods previously eaten at a particular table across multiple days); *event* memory, which requires **relating** events separated by long intervals through ordering or linking (e.g., correctly sequencing activities spread across a week); and *behavior* memory, which requires **abstracting** regularities from repeated experience (e.g., inferring where a person typically uses their phone based on accumulated observations). Together, these three types span the aggregation, relational, and inductive reasoning that genuine week-long memory entails, and that retrieval-centric evaluation cannot capture.

To construct EGOMEMREASON, we adopt a four-stage pipeline that combines **automated model-based generation with human verification**, transforming raw week-long egocentric video from EgoLife (Yang et al., 2025a) into a rigorously verified question set. We first convert each video into structured evidence through dense object-centric captioning and hierarchical event summarization with a strong MLLM. We then design task-specific query generators that extract information from the week-long video and produce candidate multiple-choice questions, each constrained to a designated query timestamp so that only past observations are accessible. Finally, all surviving candidates undergo human verification and revision using a multi-dimensional quality rubric assessing clarity, answer correctness, and option quality. Notably, annotators not only validate answers but also iteratively refine questions and distractors to eliminate ambiguity and strengthen visual grounding, resulting in 500 questions across six core challenges.

We evaluate **17** systems spanning three complementary paradigms: **8 general-purpose MLLMs** (Qwen-3-VL-8B, 30B-A3B and 32B (Bai et al., 2025), InternVL3.5-8B and 38B (Wang et al., 2025b), GPT-5 (OpenAI, 2025), Gemini-3-Flash (Google DeepMind, 2025) and Gemini-3.1-Pro (Google DeepMind, 2026)), **5 video-specific MLLMs** (LongVA (Zhang et al., 2024b), InternVideo-2.5 (Wang et al., 2025c), VideoLLaMA3 (Zhang et al., 2025a), Molmo2 (Clark et al., 2026), StreamingVLM (Xu et al., 2025)), and **4 agentic video frameworks** (SiLVR (Zhang et al., 2026), Ego-R1 (Tian et al., 2025), WorldMM (Yeo et al., 2026), AVP (Wang et al., 2025e)). Despite their model scale and pretraining, even the best model (Gemini-3-Flash) achieves only **39.6%** overall accuracy. Further analysis shows that the three memory types fail for fundamentally different reasons: entity memory is bottlenecked by fine-grained visual grounding, event memory by long-range temporal coherence, and behavior memory by abstraction over sparse repeated evidence, indicating that progress along three orthogonal axes is needed. Ablation studies further show that neither much-denser frame sampling nor auxiliary text inputs (captions, transcripts) yield consistent improvement, suggesting that the challenge cannot be resolved simply through increased visual coverage or additional textual context. Together, EGOMEMREASON and our analysis chart a path toward multimodal systems with structured, long-horizon memory that reasons beyond retrieval.

2 Related Work

Long Video Understanding (LVU). Recent work has extended video understanding from short clips to long temporal horizons. Early benchmarks (Tapaswi et al., 2016; Lei et al., 2018; Xiao et al., 2021; Wu et al., 2024a) focus on short videos with localized evidence, while newer datasets (Mangalam et al., 2023; Fu et al., 2025; Wu et al., 2024c; Yang et al., 2025a; Hu et al., 2025a; Wang et al., 2025a; Tsuchiya et al., 2026; Hummel et al., 2024) evaluate longer videos with more complex reasoning (Cheng et al., 2024; Chen et al., 2026; Yan et al., 2025; Yu et al., 2026; Chen et al., 2025; 2024). A series of methods (Yu et al., 2023; 2024; Wang et al., 2024; Zhang et al., 2024a; Tang et al., 2025; Wu et al., 2019; Fan et al., 2024; Goletto et al., 2024; He et al., 2024; Song et al., 2024) study how to address LVU from different perspectives, including external memory (Jin et al., 2025; Fan et al., 2024), agentic pipeline Wang et al. (2025e); Long et al. (2025), attention optimization (Xu et al., 2025), and so on. Recent benchmarks such as EgoLife (Yang et al., 2025a), MM-Lifelong (Chen et al., 2026), and MA-EgoQA (Kim et al., 2026b) move toward long-horizon, cross-event reasoning but differ in their treatment of memory. EgoLife, despite using ultra-long videos, relies on short-interval visual cues and lacks long-term temporal dependency. MM-Lifelong

models long-term multimodal experience with an agentic memory mechanism, but primarily evaluates retrieval over long contexts. MA-EgoQA focuses on multi-agent interaction and shared context. Overall, most LVU benchmarks remain retrieval-centric, reducing tasks to locating a few relevant segments and performing localized reasoning, without systematically studying memory mechanisms. In contrast, we explicitly study *structured memory*, where evidence is distributed over long time spans and must be incrementally accumulated before retrieval becomes meaningful.

Memory Benchmarks in Text and Multimodal Domains. Recent work evaluates long-term memory across text and multimodal settings. In text, synthetic benchmarks (Hsieh et al., 2024; Kuratov et al., 2024) offer controlled evaluation but rely on artificial signals, while task-driven (Hu et al., 2025c; Wang et al., 2025d), conversational (Maharana et al., 2024; Wu et al., 2024b), and narrative benchmarks (Wan & Ma, 2025; Kim et al., 2026a) require tracking information over extended contexts. In multimodal domains, memory is increasingly critical as information is distributed across time and modalities. Prior work in long-form and egocentric video (Yang et al., 2025a; Chen et al., 2026; Zhou et al., 2025; Perrett et al., 2025; Bärmann & Waibel, 2022; Lando et al., 2026), multimodal dialogue (Bei et al., 2026), and embodied systems (Datta et al., 2022; Savva et al., 2019; Zhang et al., 2024c) requires integrating visual, linguistic, and interaction histories over long horizons. However, these benchmarks are primarily *task-driven* and do not explicitly isolate the structure of memory required for success, leaving unclear what information is stored, how it is updated, and which memory types models rely on. EGOMEMREASON addresses this gap by explicitly decomposing long-horizon video memory into three functionally distinct types: entity, event, and behavior, which capture object-state trajectories, temporally grounded episodes, and patterns abstracted from repeated observations, respectively.

3 Benchmark Construction

Our benchmark is organized around three complementary memory types and six core challenges in total. Every question is designed to demand reasoning over accumulated temporal evidence across the long video, rather than single-clip retrieval or surface-level pattern matching. Formally, given an observed video frame sequence $V_{\leq t_q} = \{v_1, \ldots, v_{t_q}\}$ up to query time t_q and a query q, a model $f_\theta^{(m)}$ produces an answer

$$\hat{y}^{(m)} = f_\theta^{(m)}\left(q,\, \mathcal{M}_{\leq t_q}^{(m)}(V_{\leq t_q})\right), \tag{1}$$

where $m \in \{\text{entity}, \text{event}, \text{behavior}\}$ denotes the memory type, $\mathcal{M}_{\leq t_q}^{(m)}$ is the corresponding structured memory built from past observations $V_{\leq t_q}$, and $\hat{y}^{(m)}$ is the predicted answer. In the following, we present the memory design principle (§3.1), task definitions for each of the six challenges (§3.2), and the four-stage data generation pipeline together with benchmark statistics (§3.3). More detailed formulations for each task are provided in Appendix A.

3.1 EGOMEMREASON Design Principle

Cognitive science has long established that human memory is not a monolithic retrieval store but comprises qualitatively distinct systems (Kahneman et al., 1992; Tulving, 1972), each serving a different functional role: **entity memory** allows individuals to track and re-identify objects and people as they change over time, **event memory** supports the recollection of specific experienced events situated in time and place, and **behavior memory** enables the abstraction of general knowledge detached from any particular episode. These categories are complementary: each operates at a different granularity and timescale, and together they enable flexible, efficient, and robust cognitive performance. This principle has also been adopted in LLM-based agents, which increasingly incorporate structured memory modules inspired by these cognitive categories (Park et al., 2023; Zhang et al., 2025b). Recent benchmarks (Wu et al., 2024b; Maharana et al., 2024; Wan & Ma, 2025; Hu et al., 2025c) probe distinct memory competencies such as retention, retrieval, and update, but examine these capabilities in isolation without considering how different memory types interact.

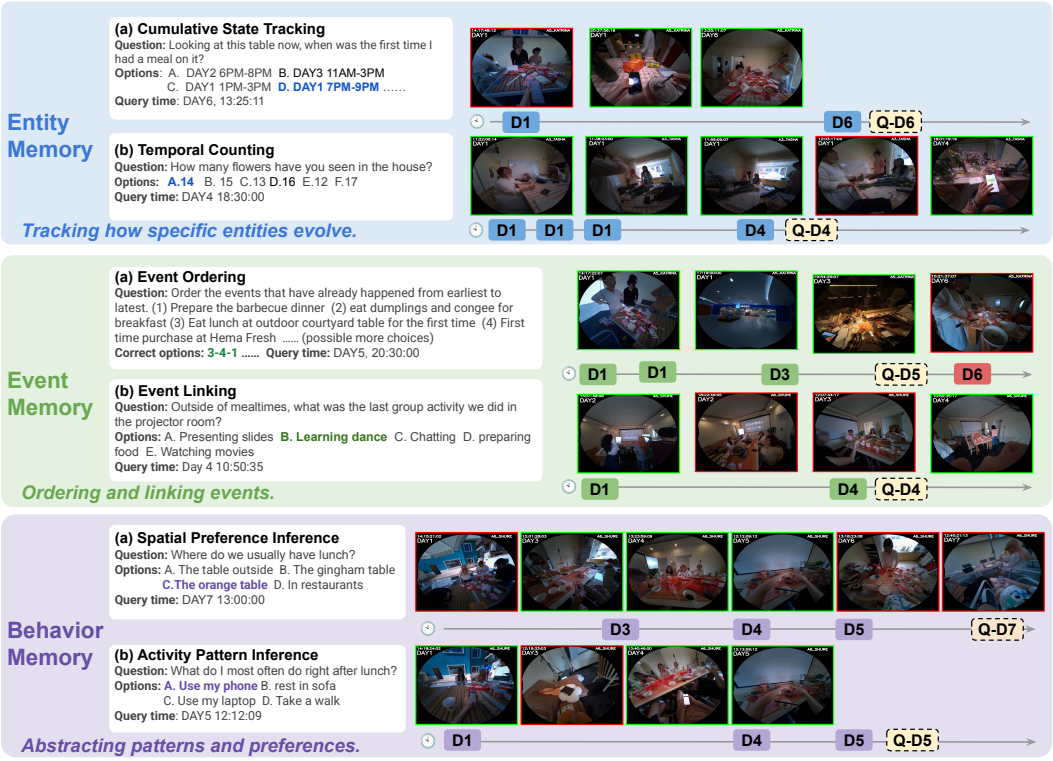

Figure 3: Overview of the six core challenges across three memory types in EGOMEMREASON. Within each example, the week-long timeline shows evidence frames sampled at different timestamps (e.g., **D1 and D4 denote evidence days, and Q-D6 indicates the query timestamp on Day 6, highlighted by a dashed box**). The timeline provides a unified temporal layout over a week and does not necessarily correspond one-to-one with the frames shown above. Green frames indicate relevant evidence; red frames indicate distractors that should not contribute to the answer — either because they are unrelated to the query, or because they occur *after* the query timestamp (e.g., D6 in Event Ordering lies beyond the Day-5 query and is therefore inadmissible evidence).

This limitation is even more pronounced in multimodal settings, where existing evaluations rarely distinguish memory types or require their integration over long temporal horizons. Evaluating these capabilities in a structured and disentangled manner is therefore critical for diagnosing model limitations in long-horizon video understanding.

3.2 EGOMEMREASON Task Definition

We formulate long-horizon video understanding as a *structured memory reasoning* problem. Building on the three memory types defined in §3.1, we operationalize each into two tasks targeting distinct reasoning demands, yielding six tasks in total (Figure 3). Each task is designed so that answering requires aggregating evidence from multiple temporally distributed observations rather than retrieving a single moment.

Entity Memory. At week-long scales, entities appear, disappear, and resurface in altered states—under different lighting, from new viewpoints, or in entirely different locations. Re-identifying them and tracking how they evolve over such intervals demands more than single-frame recognition. We evaluate this through two core capabilities:

1. *Cumulative State Tracking.* Given an entity observed at multiple points in the video, the task is to identify how its location or condition has changed across observations separated by hours or days. As shown in the cumulative state tracking example of Figure 3, the model may need to track that "a bowl" initially appears on the "kitchen counter", is later

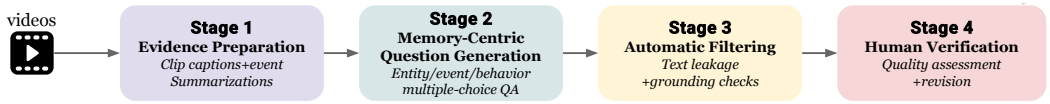

Figure 4: Overview of the benchmark construction pipeline. **Stage 1** extracts clip captions and event summaries from raw egocentric videos; **Stage 2** generates entity/event/behavior multiple-choice questions from this evidence; **Stage 3** filters text-leakage cases and verifies temporal grounding; **Stage 4** performs human quality assessment and revision.

moved to the "sink", and finally placed on the "dining table", with each observation separated by hours.

2. *Temporal Counting.* This task requires reasoning over sets of entities by counting how many distinct instances of a category have appeared across the video. Unlike static counting, the count is defined with respect to a query timestamp, such that only instances observed up to that time are considered. This requires constructing a global inventory of entity instances over time, identifying repeated occurrences under varying visual conditions, and distinguishing instances that may share similar appearance or context.

Event Memory evaluates the ability to retrieve, temporally organize, and relate discrete events from the video history. Week-long videos contain a rich stream of activities that unfold over hours or days, where later events often depend on, revisit, or modify earlier ones. We evaluate event memory through two core capabilities:

1. *Event ordering.* Given a set of events drawn from different days, the task requires arranging them in correct temporal order. This demands maintaining a structured timeline of past activities across large temporal gaps.
2. *Event Linking.* Given a set of contextual constraints (e.g., location, activity type, or time of day), the task requires identifying the relevant event matching those conditions from the previous video inputs. This requires reasoning over hours or even days, filtering candidates under multiple constraints, and selecting the correct event among visually and semantically similar alternatives.

Behavior Memory tests whether a system can abstract higher-level knowledge from repeated observations over time, going beyond individual-event recall to form generalized priors. While entity memory focuses on specific objects and event memory on specific events, behavior memory requires distilling patterns that no single observation can reveal. We evaluate behavior memory through two core capabilities:

1. *Spatial Preference Inference.* This task requires inferring recurring patterns or habitual associations from repeated observations over time, such as spatial preferences (e.g., where a person typically performs a given activity) or common activities at a location.
2. *Activity Pattern Inference.* This task requires predicting likely next states based on learned behavior patterns, such as the next location given the current one or the next activity given the current context (e.g., "Where is the person most likely to go after lunch?"). These questions test whether the system has internalized patterns from the daily routines.

Under this unified formulation, all tasks require retrieving relevant observations from long video histories, but differ in the structure of memory to be constructed: entity memory tracks persistent state trajectories, event memory organizes temporally grounded episodes, and behavior memory abstracts recurring patterns across time. This distinction enables controlled evaluation of how models store, update, and reason over long-horizon multimodal information. Together, the three memory types and their six core capabilities provide a comprehensive evaluation of the memory capacities required for long-term memory in week-long egocentric video understanding.

3.3 Benchmark Construction Pipeline and Statistics

Our benchmark is built on videos from the EgoLife dataset (Yang et al., 2025a), which provides week-long, continuous egocentric recordings across six participants in naturalistic daily routines. The multi-day, always-on nature of these

videos, capturing recurring activities, evolving object states, and extended social interactions, makes them uniquely suited for evaluating long-horizon memory. EGOMEMREASON is formulated as a multiple-choice question-answering benchmark, with each question paired with one correct answer and several semantically competitive distractors. As shown in Figure 4, we construct questions through a four-stage pipeline designed to ensure that every question is temporally grounded, visually verified, and genuinely challenging. Figure 5 summarizes the composition of EGOMEMREASON, which contains 500 questions across three memory types and six core challenges. Further details on the construction pipeline are provided in Appendix A.5.

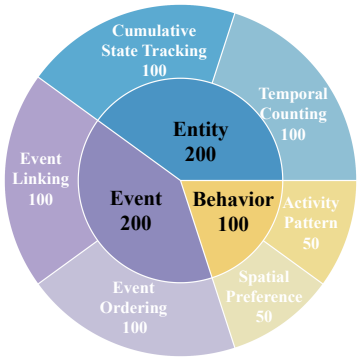

Figure 5: Dataset composition by memory type.

Stage 1: Evidence Preparation. We convert raw multi-day egocentric video into structured evidence by generating both fine-grained clip-level captions and higher-level event summaries. Videos are segmented into non-overlapping 30-second clips and processed with GPT-5 (OpenAI, 2025) to produce object-centric descriptions, tracking state changes (e.g., open/closed, on/off, and full/empty), spatial locations, and human interactions (e.g., who holds, uses, or hands off the object) as well as event-level activities and their temporal context. These dense captions are further organized into a hierarchical event structure spanning three levels of temporal granularity (30-second, 10-minute, 2-hour, and day-level), annotated with activity labels, location tags, and object references, serving as the primary signal for the query generation.

Stage 2: Query Generation. From the structured evidence, we generate candidate multiple-choice questions for each of the three memory types, each associated with a query time where only prior observations are accessible. We use GPT-5.2 (OpenAI, 2025), a different version within the same model family as the captioning stage to avoid self-reinforcing biases. For each memory type, we design a task-specific pipeline comprising three steps: (1) *statement extraction*, which identifies and aggregates relevant factual statements from the structured evidence to serve as the basis for question formulation; (2) *query generation*, which formulates questions targeting the corresponding memory capability; and (3) *distractor generation*, which produces semantically competitive incorrect options drawn from similar contexts to ensure non-trivial difficulty.

Stage 3: Automatic Filtering. We apply model-based filtering to remove trivial, ambiguous, or ungrounded questions. A blind test presents each question without visual input to three LLMs (Gemini-3.1-Pro (Google DeepMind, 2026), GPT-5 (OpenAI, 2025), Qwen-3-VL-32B (Bai et al., 2025)); questions answered correctly by a majority are discarded. We then verify that correct answers are supported by valid visual evidence before the query time, enforce a minimum temporal gap of 2 hours across supporting evidence, well beyond existing benchmarks (Fu et al., 2025; Zhou et al., 2024; Wang et al., 2025a). Details are provided in Appendix A.

Stage 4: Human Verification. All remaining candidates undergo human verification through a dedicated annotation interface that presents annotators with the query-time context video and pre-selected evidence clips alongside each question. Six annotators at the college or graduate level review each surviving question, spending approximately 20 minutes per sample to assess (1) question clarity, (2) answer correctness, and (3) option quality, and can revise or reject problematic samples. This process forms a quality-control loop in which revised questions are re-validated before inclusion, ensuring that the final benchmark is both visually grounded and human-validated. Overall, only 15% of initial candidates survive the combined model-based filtering and human verification stages, reflecting the stringent quality standards applied throughout the pipeline. We provide detailed human verification and UI in Figure 7 in the Appendix.

Method	Entity Memory		Event Memory		Behavior Memory		Overall
	Tracking	Counting	Ordering	Linking	Spatial	Activity	
Random	19.6	16.7	11.1	17.3	19.3	19.2	16.8
General MLLMs							
InternVL3.5-8B (Wang et al., 2025b)	23.0	29.0	23.0	27.0	34.0	42.0	28.0
Qwen-3-VL-8B (Bai et al., 2025)	35.0	28.0	23.0	21.0	40.0	42.0	29.6
InternVL3.5-38B (Wang et al., 2025b)	33.0	40.0	27.0	24.0	46.0	32.0	32.6
Qwen-3-VL-30B-A3B (Bai et al., 2025)	36.0	48.0	25.0	26.0	40.0	30.0	34.0
Qwen-3-VL-32B (Bai et al., 2025)	35.0	46.0	28.0	27.0	**50.0**	46.0	36.8
GPT-5 (OpenAI, 2025)	29.0	42.0	20.0	18.0	32.0	28.0	27.8
Gemini-3-Flash (Google DeepMind, 2025)	**46.0**	28.0	36.0	**44.0**	44.0	44.0	**39.6**
Gemini-3.1-Pro (Google DeepMind, 2026)	40.0	26.0	**44.0**	33.0	40.0	**48.0**	37.4
Video-specific MLLMs							
LongVA-7B (Zhang et al., 2024b)	22.0	18.0	20.0	22.0	20.0	22.0	20.6
StreamingVLM (Xu et al., 2025)	25.0	29.0	21.0	20.0	20.0	32.0	24.2
InternVideo-2.5-8B (Wang et al., 2025c)	29.0	27.0	25.0	15.0	32.0	32.0	25.6
VideoLLaMA3-8B (Zhang et al., 2025a)	23.0	31.0	27.0	32.0	38.0	36.0	30.0
Molmo2-8B (Clark et al., 2026)	36.0	**50.0**	27.0	25.0	34.0	22.0	33.2
Agentic Video Frameworks							
SiLVR (Zhang et al., 2026)	31.0	14.0	27.0	17.0	18.0	28.0	22.4
Ego-R1 (Tian et al., 2025)	30.0	18.0	23.0	18.0	48.0	32.0	25.8
WorldMM (Yeo et al., 2026)	32.0	44.0	21.0	21.0	34.0	36.0	30.6
AVP (Wang et al., 2025e)	34.0	42.0	31.0	27.0	38.0	34.0	34.0

Table 1: Main benchmark results on EGOMEMREASON. We report accuracy (%) across three memory types and six capability dimensions: Tracking (Cumulative State Tracking), Counting (Temporal Counting), Ordering (Event Ordering), Linking (Event Linking), Spatial (Spatial Preference Inference), and Activity (Activity Pattern Inference). The best result in each column is **bolded** and the second best is underlined.

4 Experimental Results

4.1 Experimental Setup

Evaluated Models. We evaluate 17 systems spanning three complementary paradigms to assess how different architectural and reasoning strategies handle memory-intensive, long-horizon tasks. This includes general-purpose MLLMs (Qwen-3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025b), GPT-5.2 (OpenAI, 2025), Gemini-3-Flash (Google DeepMind, 2025), Gemini-3.1-Pro (Google DeepMind, 2026)) covering a wide range of model scales and capabilities; video-specific MLLMs (LongVA (Zhang et al., 2024b), StreamingVLM (Xu et al., 2025)), InternVideo-2.5 (Wang et al., 2025c), VideoLLaMA3 (Zhang et al., 2025a), Molmo2 (Clark et al., 2026)) incorporating extended temporal modeling or video-centric pretraining; and agentic video frameworks (SiLVR (Zhang et al., 2026), Ego-R1 (Tian et al., 2025), WorldMM (Yeo et al., 2026)) and AVP (Wang et al., 2025e) that decompose reasoning into structured sub-tasks with retrieval or external memory.

Evaluation Metric. We report standard multiple-choice question accuracy (%) at multiple granularities: per capability (cumulative state tracking, temporal counting, event ordering, event linking, spatial preference inference, activity pattern inference), per memory type (Entity, Event, Behavior) and overall performance.

Implementation Details. For all general MLLMs and VideoLLMs, as dense 1FPS sampling from week-long video is impractical for all baselines, we uniformly sample 1024 frames from the video start to the query timestamp t_q, which are further down-scaled if necessary to fit each model's context length. For agentic video frameworks, we adopt each method's best-performing configuration reported on EgoLife (Yang et al., 2025a); when the original settings are not accessible, we default to GPT-5 (OpenAI, 2025) as the main agent. More implementation details are included in Appendix B.

4.2 Main Results

In Section 4.2, we summarize the performance of all evaluated systems on EGOMEMREASON and highlight several key findings. Across all evaluated models, performance remains low

on EGOMEMREASON, with no model achieving strong accuracy across all memory types. Results vary significantly across models and tasks, indicating that long-horizon memory reasoning remains an open challenge.

No single approach consistently leads. The best overall accuracy is Gemini-3-Flash (Google DeepMind, 2025) at 39.6%, with no model reaching 51% on any single capability. Scaling model size within families yields moderate gains: Qwen-3-VL improves by 7.2% from 8B to 32B, and InternVL3.5 gains 4.6% from 8B to 38B, suggesting that increasing parameters alone does not resolve the core challenges of long-horizon memory reasoning. Across paradigms, no category consistently outperforms the others: general MLLMs lead on Event and Behavior memory, with Gemini-3.1-Pro topping Event Ordering, Gemini-3-Flash topping Event Linking, and Qwen-3-VL-32B topping Spatial Preference Inference; video-specific MLLMs are competitive on Entity memory, where Molmo2-8B achieves the highest Temporal Counting score across all 17 systems, reflecting the benefit of pixel-level grounding pretraining; and agentic frameworks generally underperform general MLLMs, likely for two reasons: (i) most rely on caption-based intermediate representations that lose fine-grained visual cues critical for Entity Memory, and (ii) they are primarily designed for and evaluated on hour-scale videos, leaving them under-equipped for the week-long evidence aggregation required by EGOMEMREASON. Within these limits, agentic frameworks still show targeted strengths, with Ego-R1 reaching 48.0% on Spatial Preference Inference where retrieving a single dominant pattern suffices.

Different memory types expose different bottlenecks. The results show that the three memory types fail for distinct reasons, each pointing to a different missing capability rather than a shared limitation. *Entity Memory* is bottlenecked by fine-grained visual grounding combined with long-context modeling: models that rely more heavily on text-centric reasoning or training (e.g., LongVA, SiLVR) fall below 25% on Temporal Counting, while Molmo2-8B, which combines pixel-level grounding pretraining (e.g., pointing and tracking) with long-context post-training, leads all 8B models on both Cumulative State Tracking and Temporal Counting, indicating that Entity Memory benefits from both perceptual precision and the ability to retain visual evidence over extended temporal contexts. *Event Memory* is bottlenecked by long-range temporal coherence: even the strongest models get 44% accuracy on both Event Ordering and Event Linking examples. This contrasts sharply with prior egocentric benchmarks (Yang et al., 2025a; Yan et al., 2025), where models achieve much higher accuracy on single-event retrieval, suggesting that current models can locate individual events but struggle when evidence must be related across extended temporal spans, a pattern further confirmed by the sharp drop in Event accuracy as temporal certification length grows (Table 2). *Behavior Memory* is bottlenecked by long-horizon reasoning over sparse repeated evidence: even the best models stay at 50.0% on Spatial Preference Inference and 48.0% on Activity Pattern Inference. This contrasts with prior long-video benchmarks (Yan et al., 2025; Mangalam et al., 2023), where models achieve strong performance on global video summarization, suggesting that current models can summarize what they have seen but struggle to abstract recurring patterns across many sparsely distributed observations. Together, these gaps confirm that progress on long-horizon video understanding requires advances on three orthogonal axes: perceptual precision combined with long-context retention for entities, structured temporal modeling for events, and aggregation-based reasoning for behaviors, none of which are addressed by simply scaling model size or input length.

4.3 Ablation Studies and Analysis

We further conduct a series of analyses to better understand the underlying bottlenecks revealed in the main results, focusing on the roles of temporal length, visual input scaling, and auxiliary information. Unless otherwise specified, all experiments are conducted using Qwen-3-VL-8B (Bai et al., 2025). In addition, we provide further ablation studies on prompt strategies, additional analysis in Appendix C.

Effect of temporal certification. In Table 2, we analyze performance as a function of temporal certification length, defined as the total video duration one must search to locate all ground-truth evidence. Overall accuracy decreases as the temporal span increases, confirming that longer evidence spans pose a substantial challenge regardless of memory

Cert. Length (h)	<8	8–16	16–32	32+	Total
Entity	28.5	33.9	32.1	30.3	31.5
Event	–	31.1	23.0	13.5	22.0
Behavior	–	–	43.7	37.0	41.0
Overall	40.3	33.7	32.5	23.2	29.6

Table 2: Effect of temporal certification on accuracy (%) across memory types. "–" indicates bins with no examples for the corresponding memory type, as those questions occur only at longer temporal spans.

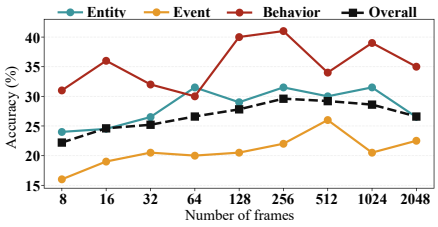

Figure 6: Effect of input frame count on accuracy across memory types.

type. The three memory types exhibit distinct degradation patterns. Event memory shows the sharpest and most monotonic decline, falling by more than half across the available ranges, indicating that event memory is the most temporally sensitive of the three. Event memory shows the sharpest decline, dropping substantially as the evidence span grows from short to medium ranges and continuing to fall at the longest spans. Behavior memory is only defined for longer spans and declines moderately as the temporal window extends. Together, these patterns show that the impact of temporal span varies substantially by memory type, indicating that benchmarks evaluating long-horizon understanding must consider per-type temporal dynamics rather than reporting a single aggregate trend.

Frame input scaling. We analyze how performance changes as the number of sampled input frames increases, as shown in Figure 6. Entity memory improves steadily with more frames before saturating around 256, indicating that visual coverage helps up to a point but cannot compensate for models' limited capacity to extract fine-grained object states from individual frames. Event memory is the least responsive to frame scaling, which suggests that the bottleneck is not visual coverage but the inability to maintain long-range temporal coherence when ordering and linking events across days. Behavior memory benefits most from denser sampling but remains highly unstable across budgets, reflecting its dependence on capturing recurring patterns, a signal that is easily disrupted when additional frames introduce conflicting or off-routine observations. Overall, no single frame budget is optimal across different memory types, indicating that scaling frame count alone cannot address long-horizon memory.

Quantitative error analysis. To better understand the limitations of current MLLMs, following existing work (Cheng et al., 2025), we randomly sampled 100 benchmark examples and manually inspected the failure cases. Our analysis on Gemini-3-Flash reveals four common failure patterns: (1) extracting incorrect visual details from the video (28%), (2) missing important visual information, especially in long-range reasoning scenarios (32%), (3) perceiving the correct visual details but making logical mistakes during reasoning (32%), and (4) producing incorrect predictions despite correct reasoning processes (8%).

5 Conclusion

We introduced EGOMEMREASON, a benchmark for evaluating long-horizon memory reasoning in week-long egocentric videos, decomposing memory into three complementary types (entity, event, and behavior) across six core challenges that require multi-hop reasoning over temporally distributed evidence. Evaluating 17 systems reveals that long-horizon memory remains a substantial open challenge: the three memory types fail for fundamentally different reasons. Entity memory is bottlenecked by fine-grained visual grounding over long time period, event memory by long-range temporal coherence, and behavior memory by abstraction over sparse repeated evidence. We further show through ablations on temporal length, frame input, and auxiliary information that none of these bottlenecks can be addressed simply by scaling input or context. We believe EGOMEMREASON serves as a rigorous diagnostic framework for guiding future research toward models capable of genuine long-horizon memory reasoning.

Acknowledgment

We would like to thank David Wan, Nithin Sivakumaran, and Fengli Wu for their help in the human annotation process. This work was supported by ONR Grant N00014-23-1-2356, ARO Award W911NF2110220, DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, Laboratory for Analytic Sciences via NC State University, National Institutes of Health Award 1R01HD111074-01, and Sony Focused Research award. The views contained in this article are those of the authors and not of the funding agency.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 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 (CVPR) Workshops*, pp. 1560–1568, June 2022.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515*, 2026.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024.

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, Baoqi Pei, Shihao Wang, et al. Towards multimodal lifelong understanding: A dataset and agentic baseline. *arXiv preprint arXiv:2603.05484*, 2026.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George

Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19119–19128, 2022.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

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

Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*, 2024.

Google DeepMind. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/, 2025.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026. Accessed: 2026-04-25.

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, 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 (CVPR)*, pp. 19383–19400, June 2024.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13504–13514, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025a.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *arXiv preprint arXiv:2505.22657*, 2025b.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025c.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. The reviewing of object files: object-specific integration of information. *Cognitive psychology*, 24 2:175–219, 1992. URL https://api.semanticscholar.org/CorpusID:2688060.

Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha, and Minjoon Seo. Can large language models keep up? benchmarking online adaptation to continual knowledge streams. *arXiv preprint arXiv:2603.07392*, 2026a.

Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. Ma-egoqa: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827*, 2026b.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Giuseppe Lando, Rosario Forte, Giovanni Maria Farinella, and Antonino Furnari. How far can off-the-shelf multimodal large language models go in online episodic memory question answering? In Emanuele Rodolà, Fabio Galasso, and Iacopo Masi (eds.), *Image Analysis and Processing – ICIAP 2025*, pp. 499–511, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-10192-1.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arxiv*, 2024.

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.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. Minerva: Evaluating complex video reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23968–23978, October 2025.

OpenAI. GPT-5 system card, 2025.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23901–23913, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Borislav Ivanovic, Julian Straub, Jia Liu, Vladlen Koltun, and Jitendra Malik. Habitat: A platform for embodied ai research. In *ICCV*, 2019.

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.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4631–4640, 2016.

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, 2025. URL https://arxiv.org/abs/2506.13654.

Fumihiko Tsuchiya, Taiki Miyanishi, Mahiro Ukai, Nakamasa Inoue, Shuhei Kurita, Yusuke Iwasawa, and Yutaka Matsuo. Ec-bench: Enumeration and counting benchmark for ultra-long videos, 2026. URL https://arxiv.org/abs/2603.29943.

Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory*, pp. 381–403. Academic Press, New York, 1972.

Luanbo Wan and Weizhi Ma. Storybench: A dynamic benchmark for evaluating long-term memory with multi turns. *arXiv preprint arXiv:2506.13356*, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

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*, 2025b.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025c.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025d.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025e.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024a.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 284–293, 2019.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. 2024b. URL https://arxiv.org/abs/2410.10813.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024c.

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.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, Changzhi Sun, et al. Teleego: Benchmarking egocentric ai assistants in the wild. *arXiv preprint arXiv:2510.23981*, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17294–17303, 2025c.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Shoubin Yu, Lei Shu, Antoine Yang, Yao Fu, Srinivas Sunkara, Maria Wang, Jindong Chen, Mohit Bansal, and Boqing Gong. Ego2web: A web agent benchmark grounded in egocentric videos. *arXiv preprint arXiv:2603.22529*, 2026.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Guanzheng Chen Yuqian Yuan, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, November 2024a.

Ce Zhang, Yan-Bo Lin, Ziyang Wang, Mohit Bansal, and Gedas Bertasius. SiLVR: A simple language-based video reasoning framework. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=mQZbh9Zlbw.

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*, 2024b. URL https://arxiv.org/abs/2406.16852.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024c.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *ACM Transactions on Information Systems*, 2025b. doi: 10.1145/3748302.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2024. URL https://api.semanticscholar.org/CorpusID:270286192.

Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio W. Mayol-Cuevas, Fan Zhang, Weizhe Lin, and Junxiao Shen. X-LeBench: A benchmark for extremely long egocentric video understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15206–15222, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.822. URL https://aclanthology.org/2025.findings-emnlp.822/.

Appendix

In this appendix, we first describe the full data construction pipeline (§A). We then present additional implementation details (§B) and additional analysis (§C). Finally, we discuss limitations and directions for future work (§D).

A Data Construction Details

Our benchmark is built on videos from the EgoLife dataset (Yang et al., 2025a), which provides ultra-long, continuous egocentric recordings spanning multiple days across six participants engaged in naturalistic daily routines. This multi-day, always-on nature of the video makes it uniquely suited for evaluating long-horizon memory: the recordings capture rich temporal dynamics, including recurring activities, evolving object states, and extended social interactions that unfold across days rather than minutes.

As shown in previous Figure 4, we construct questions through a four-stage pipeline designed to ensure that every retained item is temporally grounded, visually verified, and genuinely challenging.

A.1 Stage 1: Evidence Preparation

The first stage converts raw multi-day egocentric video into structured textual evidence that supports downstream question generation. We segment each participant's recording into 30-second clips and caption them with a VLM following a structured, object-centric rubric. These clip-level captions are then aggregated into event-level summaries annotated with activity labels, location tags, and object references. The resulting dual-granularity representation, consisting of fine-grained clip timelines paired with coarser event scaffolds, serves as the primary input for question generation.

Clip-level Captioning. We segment each participant's recording into non-overlapping 30-second clips and sample frames within each clip at 1 FPS. A VLM (GPT-5) receives the sampled frames and produces object-centric state-tracking annotations following a structured rubric. The rubric directs the model to attend to five dimensions in order: presence, including appearance, disappearance, addition, and removal; attribute or status changes such as open/closed, on/off, full/empty, or clean/dirty; spatial location and movement; interaction with people, covering who holds, uses, or hands off each object; and count or accumulation over time. For each clip, the model produces a dense caption that records, where applicable, each object's current status, location, interaction context, what changed relative to earlier in the clip, and the specific evidence frames. This object-centric design grounds the captions in observable states rather than narrative summaries, providing the fine-grained temporal evidence needed for downstream entity-tracking and episodic-memory questions.

Hierarchical Event Summarization. Beyond clip-level captions, we summarize each participant's recording at multiple temporal scales along a fixed pyramid: 30-second clips, 10-minute windows, 2-hour windows, and full-day summaries. Each level is produced by prompting GPT-5 to summarize the captions from the level below into a progressively coarser description that retains activity, location, and object references. The resulting representation provides complementary views of the same video, ranging from fine-grained clip-level evidence for precise temporal grounding to day-level scaffolds that capture the overall arc of a participant's routine. This multi-scale representation serves as the primary input for all subsequent question generation stages.

A.2 Stage 2: Query Generation

From the structured evidence, we generate candidate multiple-choice questions for each of the three memory types, each associated with a query time at which only prior observations are accessible. We use GPT-5.2 (OpenAI, 2025), a different version within the same model

family as the captioning stage to avoid self-reinforcing biases. For each memory type, we apply a task-specific pipeline comprising three steps. *Statement extraction* identifies and aggregates relevant factual statements from the structured evidence to serve as the basis for question formulation. We guide the extractor with memory-type-specific in-context examples: entity-memory prompts demonstrate statements about object states and counts, event-memory prompts demonstrate temporally anchored activities, and behavior-memory prompts demonstrate recurring patterns and cross-event transitions. *Query generation* formulates a multiple-choice question targeting the relevant memory capability from each extracted statement, together with the ground-truth answer. *Distractor generation* produces competitive incorrect options for each question. Rather than sampling distractors from unrelated content, we condition the generator on other visual information from the same participant's recording, specifically statements that share salient attributes with the ground truth (such as the same object category, activity, location, or a nearby temporal window) but diverge in the queried dimension. This keeps distractors plausible against the participant's actual memory trace and rules out shortcuts based on global implausibility.

A.3 Stage 3: Automatic Filtering

Raw candidates pass through several filtering stages to ensure that every retained question is genuinely challenging, visually grounded, and non-redundant.

Blind-test filtering. We first reject samples that can be answered without video evidence. A text-only leakage test attempts to answer each question using only the query and options, flagging samples where the correct answer is recoverable from surface cues. We target five such cues: explicit answer leakage in the query text; temporal markers (e.g., "Day 3" or "10:30 AM") that trivially disambiguate options; lexical shortcut patterns; high token overlap between the query and the correct option relative to the distractors; and common-sense priors that make one option obviously dominant. For episodic memory questions, we additionally apply a stricter model-based variant: an LLM receives only the question and the options under multiple permutations of the option order, and samples whose accuracy across permutations exceeds a fixed threshold are rejected.

Grounded evidence verification. Every retained question is verified against the original video to ensure correctness. We first confirm that each referenced evidence clip exists in the recording and contains enough sampled frames for meaningful visual inspection. We then check that the captions associated with each evidence clip are consistent with the question's expected answer, so that the claimed visual evidence actually supports the correct option. All evidence clips must fall strictly before the question's query timestamp, preserving the constraint that only past observations are accessible. For event-ordering questions, we additionally verify that the temporal order of the referenced clips matches the ground-truth sequence.

A.4 Stage 4: Human Verification

After automatic filtering, all surviving candidates enter a human verification stage through a purpose-built annotation interface (Figure 7).

Annotation protocol. For each candidate question, the annotator is presented with: (i) the query-time context video, showing what the camera wearer observes at the moment the question is posed; (ii) one or more expandable evidence clips drawn from earlier in the timeline, which the pipeline identified as supporting the correct answer; and (iii) the full set of multiple-choice options with the machine-generated ground-truth label highlighted. A browsable list of related clips is also available so that annotators can independently verify evidence beyond the pre-selected set. Annotators follow five guidelines: watch the query context video and at least one relevant support clip; select the best answer based on visual evidence rather than metadata; rate query quality and revise unclear or awkward phrasing; flag option-set issues such as weak, duplicated, or ambiguous distractors; and record an overall decision indicating whether the question can be kept as-is.

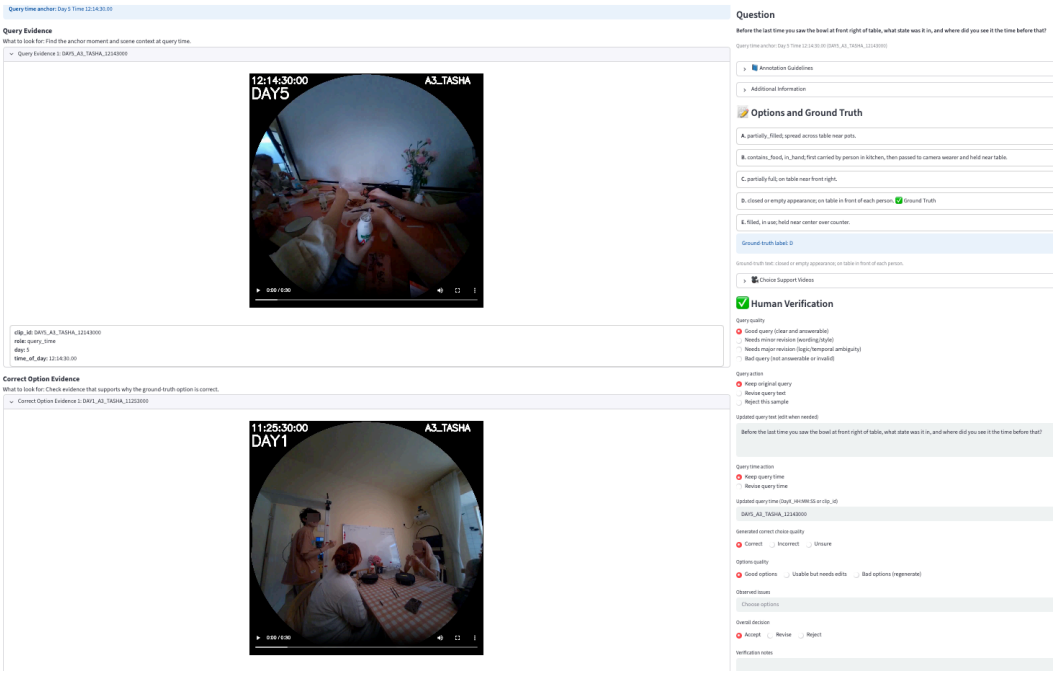

Figure 7: Human verification interface. **Left:** the annotator reviews the query-time context video and expandable evidence clips alongside the question and option set. **Right:** multi-dimensional quality assessment panel with structured ratings for query quality, correct-choice quality, option quality, and an overall accept/revise/reject decision.

Multi-dimensional quality assessment. Rather than a single accept or reject label, the interface collects fine-grained judgments along four axes. *Query quality* is rated on a four-point scale: good (clear and answerable), needs minor revision (wording or style), needs major revision (logic or temporal ambiguity), or bad (not answerable or invalid). *Correct-choice quality* is assessed as correct, incorrect, or unsure, allowing annotators to flag cases where the machine-generated ground truth does not match the visual evidence. *Option quality* is rated as good, usable but needs edits, or bad (regenerate), capturing distractor-level issues that may not affect the correct answer itself. An *observed issues* field records specific failure modes such as temporal inconsistency, duplicate options, or answer leakage.

Revision and decision. Based on these assessments, annotators select a query action: keep the original text, revise it with an editable text field for in-place rewording, or reject the sample entirely. A final overall decision of **Accept**, **Revise**, or **Reject** is recorded alongside free-form verification notes for borderline cases. For revised samples, the annotator-edited query replaces the machine-generated text, and the updated question undergoes a second round of automatic verification (blind-test and evidence checks) before inclusion in the final benchmark.

Quality loop. The full annotation cycle operates as a closed loop: (1) generate the candidate set, (2) run strict automatic filtering, (3) human annotate with multi-dimensional assessment, (4) apply targeted rewrites on revised samples, and (5) re-verify and freeze the final release. We report human acceptance, revision, and rejection rates as part of the benchmark's quality diagnostics.

Tier-2 expert audit. Candidates that pass the first-tier annotation undergo a second-tier audit conducted by the authors. Unlike the first tier, which is scoped to the query-time context and the pipeline-selected evidence clips, the second-tier auditor reviews the participant's full multi-day recording before judging each question. This wider context lets the auditor assess the question against the participant's complete activity history rather than a pre-filtered slice, and surface failure modes that are invisible at the clip level: alternative valid answers supported by evidence the pipeline did not flag, distractors that are in fact true of the participant at some other point in the recording, temporal-constraint violations in which a question's correct answer becomes determinable only after the query timestamp, and ambiguities that arise when the same object, location, or activity recurs across days. For each question, the auditor independently re-answers without consulting the proposed ground-truth label, then compares the two; mismatches trigger adjudication and either revision or rejection. The auditor also re-examines the option set for redundancy with the broader recording and rewrites distractors that prove non-competitive once the full context is known. Only questions cleared at this tier are admitted to the final release.

A.5 Benchmark Details and Statistics

We detail the models and key configurations used in each stage of the construction pipeline. In Stage 1 (Evidence Preparation), we use GPT-5 (OpenAI, 2025) to generate dense clip-level captions and hierarchical event summaries from the raw egocentric video. In Stage 2 (Query Generation), we use GPT-5.2 (OpenAI, 2025) to produce candidate multiple-choice questions for each memory type, conditioned on the structured evidence from Stage 1. In Stage 3 (Automatic Filtering), we employ three models, Gemini-3.1-Pro (Google DeepMind, 2025), GPT-5.2 (OpenAI, 2025), and Qwen-3-VL-32B (Bai et al., 2025), to perform two rounds of filtering: first, a text-only leakage test that removes questions where a majority of the three models can answer correctly without visual input, and second, a quality check that verifies answer correctness, distractor plausibility, and multi-timestamp grounding. In Stage 4 (Human Verification), six annotators at the college or graduate level review each surviving question, spending approximately 20 minutes per sample to assess question clarity, answer correctness, and option quality. Only 15% of candidates pass this final stage, reflecting the stringent quality standards applied throughout the pipeline.

B Additional Implementation Details

Following existing works (Yang et al., 2025a), in EGOMEMREASON, each question is associated with a designated query timestamp t_q. Specifically, only video content observed before t_q is made accessible to the model, ensuring that no future information can be used to answer the question. Since the videos span multiple days, t_q can range from Day 1 to Day 7, and the temporal gap between the earliest required evidence and t_q routinely exceeds one full day. This constraint is enforced consistently across all evaluated systems. For all MLLM-based methods, we run experiments for the highest possible frame settings. For Qwen-3-VL-8B(Bai et al., 2025) and Molmo2-8B(Clark et al., 2026) which we conduct analysis on, we report the best overall performance setting in the main table. For agentic approaches, we follow the WorldMM(Yeo et al., 2026) to run the caption model with 1FPS sampling on 30s clips.

C Additional Experimental Results

C.1 Detailed Error Analysis

Quantitative Error Analysis. To better understand the limitations of current MLLMs, we manually inspected 100 benchmark examples and analyzed the common failure patterns. The analysis is based on Gemini-3-Flash responses. Inspired by Video-Holmes (Cheng et al., 2025), we group the failure cases into the following four categories:

1. *Visual Perception Error (VPE)* happens when the model looks at the video but extracts wrong visual details. The necessary visual information for answering the given question is captured, but the model hallucinates on the visual information.
2. *Visual Omission Error (VOE)* happens when the model simply misses important visual information. This usually happens because the questions require long-range reasoning over the entire video but the current models are restricted by the limited context window.
3. *Reasoning Error (RE)*) happens when the model sees the right visual details but then makes logical mistakes when processing the visual information. For example, the model correctly tracks the location of the target object across multiple timestamps but synthesizes the temporal information incorrectly.
4. *Think-Right-Answer-Wrong Error (TRAW)*) happens when the reasoning process is correct and aligns with human reasoning for answering the question, but the model still produces an incorrect prediction.

We find that reasoning Error and Visual Omission Error are the two dominant failure modes, accounting for 32% of errors, followed by Visual Perception Error at 28%. The relatively small proportion of Think-Right-Answer-Wrong errors (8%) suggests that when models reason correctly, they usually arrive at the correct answer. The near-equal split among the three major error types indicates that improving long-horizon video understanding requires advances on multiple fronts: more faithful visual perception, broader temporal coverage to reduce evidence omission, and stronger temporal reasoning capabilities to correctly synthesize information across distant observations.

C.2 Robustness Analysis

Robustness to Filtering Model. To assess sensitivity to the choice of filtering model, we randomly sample 100 final benchmark items and rerun the blind filtering stage on their corresponding pre-human-verification candidates, replacing Qwen-3-VL-32B with InternVL3.5-38B while keeping GPT-5 and Gemini-3.1-Pro fixed. Under the alternative filtering ensemble, 93 of the 100 candidates are retained. On the corresponding final benchmark items, Qwen-3-VL-32B and Gemini-3-Flash achieve 37.6% and 39.8% accuracy, respectively, compared with 36.8% and 39.6% on the full benchmark. These results indicate that the overall performance trend is stable under the alternative filtering configuration.

Cross-Model Consistency. To examine whether the effect of temporal span is specific to the 8B-scale model, we further evaluate Qwen-3-VL-32B and Gemini-3-Flash across the same temporal-certification ranges. As shown in Table 3, all three models exhibit a consistent degradation as the required temporal span increases, with the lowest performance on the longest-range questions. This suggests that the effect of temporal span persists across different model scales and families.

Model	<8h	8–16h	16–32h	32h+	Overall
Qwen-3-VL-8B	40.3	33.7	32.5	23.2	29.6
Qwen-3-VL-32B	46.1	40.5	38.8	31.0	36.8
Gemini-3-Flash	48.1	46.0	41.0	32.7	39.6

Table 3: Cross-model performance across temporal-certification ranges. Performance generally decreases as the temporal span required to locate all supporting evidence increases.

Robustness to Input Scaling. We further examine the effect of input frame budget across models of different scales. As shown in Table 4, increasing the number of input frames generally improves performance for Qwen-3-VL-8B, Qwen-3-VL-32B, and Gemini-3-Flash, but the gains diminish at larger budgets and can slightly decrease with further scaling. This consistent pattern suggests that simply increasing visual input coverage does not fully resolve the challenges of long-horizon video understanding.

Model	32	64	128	256	512	1024
Qwen-3-VL-8B	26.0	27.4	28.5	29.6	28.4	28.3
Qwen-3-VL-32B	30.4	32.0	35.2	35.6	36.8	36.2
Gemini-3-Flash	32.6	33.4	36.2	39.0	39.6	37.8

Table 4: Cross-model performance (%) under different input frame budgets. Increasing the number of frames generally improves performance up to an intermediate or high budget, after which the gains plateau or slightly decrease.

Trans.	Caption	Entity	Event	Behavior	All
✗	✗	**31.5**	22.0	41.0	29.6
✓	✗	29.0	**23.0**	**46.0**	**30.0**
✗	✓	29.5	21.0	45.0	29.2
✓	✓	**31.5**	19.0	45.0	29.2

Table 5: Effect of auxiliary text inputs (transcript, captions) on accuracy (%).

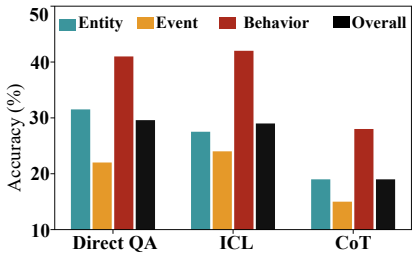

Figure 8: Effect of different prompt strategies, including direct QA, CoT prompting, and in-context learning.

C.3 Additional Benchmark Analysis

Auxiliary input information. We study the impact of additional textual inputs (transcripts and captions). As shown in Table 5, auxiliary text affects each memory type differently. Entity Memory is largely insensitive, as precise object states must be read from frames; Event Memory is consistently hurt by captions, likely because dense per-clip captions fragment the cross-clip temporal continuity needed for ordering; and Behavior Memory is the only type that benefits, with transcripts yielding the largest gain, as speech-based signals provide complementary routine and social context. Despite these per-type differences, the overall gains are marginal: only transcripts yield a small improvement (0.4%), while captions provide no benefit. This suggests that additional textual signals alone are insufficient to address the challenges of long-horizon memory reasoning.

Prompting strategy. In Figure 8, we compare direct QA, in-context learning (ICL), and chain-of-thought (CoT) prompting. The three strategies expose where the difficulty of long-horizon memory tasks actually lies. CoT prompting degrades performance substantially across all memory types, cutting overall accuracy by roughly a third and reducing every per-type score, indicating that explicit step-by-step reasoning does not help on memory-intensive tasks and instead amplifies errors that compound over long temporal contexts. ICL yields performance comparable to direct QA overall, with a small gain on Event memory offset by a modest drop on Entity memory, suggesting that the task format is already well-specified through instructions alone and that additional in-context examples offer little leverage when the underlying challenge is recall rather than format. Overall, direct QA remains near-optimal, suggesting that more elaborate prompting alone is insufficient to address the dominant challenges in long-horizon video understanding. This suggests that future improvements may benefit more from enhancing how models encode and access long-horizon visual information than from more sophisticated prompting techniques.

Oracle-Evidence Analysis. To disentangle evidence localization from reasoning over the retrieved evidence, we additionally evaluate Gemini-3-Flash by providing only the ground-truth evidence clips required for each question. Its accuracy increases from 39.6% under the full-video setting to 67.2% with oracle evidence. This substantial improvement indicates that locating the relevant evidence within week-long video is a major source of difficulty, while the remaining gap shows that reasoning over the retrieved multi-event evidence remains challenging.

Human Verification Outcomes. Human review substantially modifies the automatically generated candidates: 41.8% are directly rejected, 6.2% are refined for wording, 29.4% undergo substantial revisions to both the question and answer, 11.6% require updates to the correct answer or option set, and only 11.0% are retained unchanged. Among candidates retained after human verification, 70.4% have their machine-generated answer supervision substantially revised. These statistics indicate that the final benchmark is extensively human-validated rather than directly inheriting model-generated labels.

Query-Only Evaluation. To assess whether the benchmark can be solved through linguistic priors alone, we evaluate Qwen-3-VL-32B and Gemini-3-Flash using only the question and answer options, without any video input. They achieve 18.8% and 20.4% accuracy, respectively, close to the 16.8% random baseline and substantially below their full-video performance of 36.8% and 39.6%. This suggests that visual evidence is essential for solving the benchmark.

Agentic Failure Analysis. To understand why agentic video frameworks remain limited on EGOMEMREASON, we manually analyze 100 failed SiLVR examples by inspecting its retrieved textual memory and reasoning traces. Among these failures, 35% are caused by retrieval misses, 30% by perception errors, 32% by reasoning errors, and 3% by answer mismatches. These results show that agentic systems fail at multiple stages, suggesting that structured external memory alone does not eliminate errors in evidence access, visual grounding, and downstream temporal reasoning.

Caption Quality. We additionally compare our dense captions with the original EgoLife annotations on 100 randomly sampled clips. Our captions contain 190.7 words, 10.2 distinct objects/entities, and 15.7 action/event verbs on average, compared with 10.8, 1.4, and 1.2, respectively, in the original annotations. This reflects the substantially finer object-, action-, and state-level detail used for constructing EGOMEMREASON. We will release the intermediate captions, hierarchical summaries, and associated evidence annotations together with the benchmark.

D Limitation

Our benchmark is constructed from the week-long recordings in EgoLife, which contain six participants and therefore cover a relatively limited range of individuals, environments, and Behavior routines. This may particularly limit the diversity of Behavior Memory questions. Our goal is to evaluate long-horizon evidence accumulation and reasoning rather than population-level Behavior generalization; extending EGOMEMREASON to broader participants and environments is an important direction for future work. Also, while we carefully design tasks to require multi-hop reasoning over temporally distributed evidence, the benchmark is still constructed through controlled generation and filtering procedures. As such, it may not cover the full diversity of real-world long-horizon scenarios. Finally, our evaluation relies on sampled frames and optional auxiliary inputs (e.g., transcripts and captions), which may not fully capture all temporal dynamics in raw video streams. Future work could explore richer video representations and streaming-based evaluation settings.