APM-BENCH: BENCHMARKING CROSS-SESSION PERSISTENT MEMORY FOR REAL-WORLD EGOCENTRIC STREAMING VIDEO ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

To serve as real-world personal assistants, streaming video models need persistent memory that retains past experiences for later use. Yet existing streaming benchmarks and methods often focus on individual continuous videos or short clips, overlooking that real-world interactions are often intermittent and require memory to persist across interruptions. To fill this gap, we introduce APM-Bench, which reformulates real-world streaming interaction as multi-session life trajectories. It contains 549 sessions, 104 trajectories, and 2,719 candidates, spanning both objective and open-ended questions. Each session is a video with fine-grained annotations, and sessions within a trajectory revolve around related activities. Models then use persistent memory to answer questions about past sessions and provide proactive responses while maintaining real-time interaction. This raises challenges: persistent memory must be storable, selectively retain information, be injected at the right time, and remain efficient. Moreover, finite storage may leave required evidence unavailable, so assistants should recognize missing evidence. Therefore, we systematically evaluate general video models under different memory protocols and diverse specialized streaming memory systems, and test whether models acknowledge insufficient evidence. Our evaluation reveals a clear utility–latency–storage trade-off: existing methods still struggle to simultaneously achieve reliable long-term recall, low overhead, and effective proactive assistance across sessions. APM-Bench provides a comprehensive testbed for developing and comparing persistent memory systems under realistic streaming conditions. We hope it encourages future work that jointly considers utility, latency, and storage toward more practical persistent memory for real-world streaming assistants.

1 INTRODUCTION

Streaming video models increasingly support continuous perception, real-time interaction, and proactive assistance (Yao et al., 2026; Ant Group, 2026), showing their potential as personal assistants; memory is key to making such assistants truly personal by retaining and reusing user-specific experience over time. However, most existing benchmarks and methods study memory within a single continuous video, typically over a limited time span. In the real world, interactions are intermittent: users may turn off smart glasses and resume using the assistant hours or days later. The assistant must therefore retain and use relevant visual evidence from earlier interactions to answer later questions and provide proactive assistance.

This calls for **persistent memory**: a storable record of past experience that remains available after an interaction ends and can be reused in later interactions. For real-world assistants, ***such memory must support later tasks while keeping storage and response latency manageable***. Yet existing evaluations rarely assess memory utility together with these deployment costs. Streaming video benchmarks evaluate understanding within individual videos (Li et al., 2025; Lin et al., 2024) and extend interaction to longer continuous streams (Zhang et al., 2026b). At much longer timescales, benchmarks assess streaming episodic memory (Forte et al., 2026) and long-term proactive service (Sitong et al., 2026), while proactive interaction benchmarks evaluate when models should respond and what assistance they should provide (Zhang et al., 2025c; Ran et al., 2026; Zhao et al., 2026). Overall, previous evaluations do not yet provide a clear picture of how persistent memory

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.

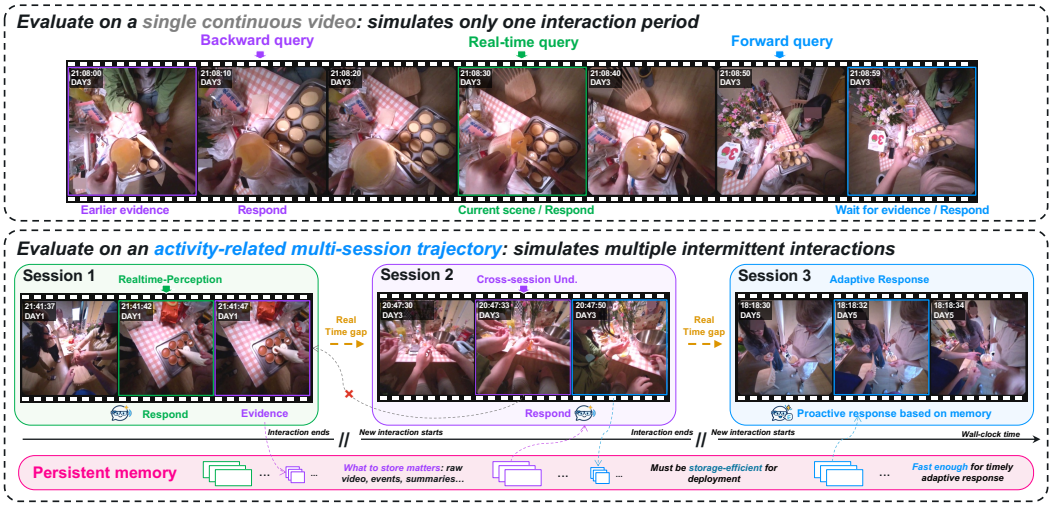

Figure 1: In a streaming setting, once an interaction ends, the model can no longer directly access its visual stream because real-world interactions are not replayed; the assistant therefore needs persistent memory to retain prior experience. Across sessions, the assistant updates and reuses persistent memory for cross-session understanding and proactive assistance while continuing real-time perception. The example shows repeated collaborative dessert-making across multiple sessions.

supports retrospective understanding and proactive assistance across temporally separated interactions while balancing utility, storage, and response latency.

To simulate such intermittent real-world use, we introduce **APM-Bench**, which organizes egocentric experience into multi-session life trajectories, as shown in Figure 1. Sessions within a trajectory contain related activities and preserve their temporal order and time gaps. Each session includes fine-grained annotations of evidence time intervals, query times, reference proactive responses, etc. This organization naturally reduces the total video duration relative to a complete life log and enables us to compare memory effectiveness, storage costs, and response latency. APM-Bench evaluates three capabilities through 12 tasks: *Cross-session Understanding* measures how well models use memory to answer questions about past sessions; *Real-time Perception* assesses models' ability to understand the current visual scene and how memory affects this ability; and *Adaptive Response* evaluates whether models provide appropriate help when needed and remain silent otherwise, with the help of memory.

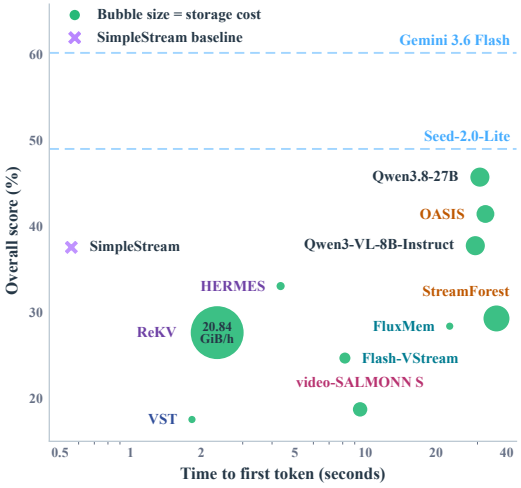

Figure 2: Utility-latency-storage trade-off across evaluated methods. Bubble size represents storage cost per hour. All general video models are evaluated under *Raw Video as Memory*.

APM-Bench highlights three challenges in building effective persistent memory for real-world assistants. First, ***what to store***: to make past experience available in later sessions, the simplest strategy is to retain raw video, while compact persistent memory must selectively preserve information and represent it in forms such as visual tokens, structured events, or model parameters. Second, ***when to use***: persistent memory is not necessary for every interaction, and irrelevant historical information can interfere with current perception (Shen et al., 2026; Ge et al., 2026). Third, ***efficiency***: the latency and storage costs introduced by persistent memory must remain manageable. More-

Table 1: Comparison of streaming video benchmarks. Most prior benchmarks evaluate models on a single continuous video; APM-Bench evaluates interactions across sessions separated by interruptions, simulating intermittent real-world use and testing whether memory from earlier sessions remains useful. It also tests whether models recognize when required historical evidence is unavailable. RTP: real-time perception; RET: retrospective tasks; PRO: proactive response; OE: open-ended candidates; OBJ: objective candidates; SE: storage-efficiency evaluation; MS: multi-session evaluation across related activities; EA: Evidence Availability-Aware evaluation.

Benchmark	RTP	RET	PRO	OE	OBJ	SE	MS	EA
StreamingBench (Lin et al., 2024)	✓	✓	✓	✗	✓	✗	✗	✗
OVO-Bench (Li et al., 2025)	✓	✓	✓	✗	✓	✗	✗	✗
PhoStream (Lu et al., 2026)	✓	✓	✓	✓	✗	✗	✗	✗
EgoStream (Forte et al., 2026)	✓	✓	✗	✗	✓	✓	✗	✗
EgoServe (Sitong et al., 2026)	✗	✗	✓	✓	✗	✗	✗	✗
StreamArena (Zhang et al., 2026b)	✓	✓	✓	✓	✗	✗	✗	✗
APM-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

over, persistent memory cannot retain an unbounded visual history under finite storage, so assistants should acknowledge when relevant evidence is unavailable rather than fabricating an answer.

We evaluate general video models by replaying stored videos at query time (*video as memory*) or session summaries supplied with the query (*text as memory*), alongside specialized memory systems. Video as memory preserves visual details but requires more storage, increases latency, and often weakens real-time perception; specialized systems vary widely in efficiency, yet most deliver weak service quality, as illustrated in Figure 2. Tests with unavailable historical evidence further show that even the strongest general video model struggles to acknowledge insufficient evidence.

Our contributions are as follows:

- We introduce APM-Bench, with 2,719 human-refined candidates spanning objective and open-ended questions, averaging 69 minutes of video per trajectory.
- APM-Bench highlights challenges in making persistent memory practical and jointly evaluates cross-session understanding, real-time perception, and adaptive response within each trajectory, with controlled tests of responses to unavailable historical evidence.
- We systematically analyze memory representations and systems across task performance, storage, and latency, revealing the strengths and limitations of existing systems.

We hope to encourage research on streaming persistent memory that jointly considers utility, latency, and storage, and to support more usable memory systems for real-world assistants.

2 Related Work

Streaming Video Benchmarks. Streaming video benchmarks require models to respond as video arrives, using only what they have seen. StreamingBench (Lin et al., 2024) and OVO-Bench (Li et al., 2025) evaluate online understanding; RTV-Bench (Xun et al., 2025) examines continuous perception and reasoning. For interaction, RIVER (Shi et al., 2026) and EgoSAT (Lei et al., 2026) combine retrospective, current, and prospective tasks; PhoStream (Lu et al., 2026) studies mobile scenarios; StreamArena (Zhang et al., 2026b) studies hour-scale interaction. Memory and proactive service are also evaluated: EgoStream (Forte et al., 2026) tests episodic recall across horizons, and EgoServe (Sitong et al., 2026) tests long-term proactive service. ESTP-Bench (Zhang et al., 2025c), EgoPro-Bench (Ran et al., 2026), and OmniPro (Zhao et al., 2026) further assess proactive response timing and content. These benchmarks move streaming evaluation toward real assistance, but mostly use single continuous videos, as shown in Table 1. APM-Bench asks how retained memory supports later interactions across temporally separated sessions, and at what storage and latency cost.

Streaming Memory Systems. Streaming video memory methods have been extensively studied to determine what to retain from a growing visual stream. Visual representation methods compress



Figure 3: **Examples of intra-session and inter-session streaming tasks.** Each session follows its original continuous wall-clock timeline. Across the 12 tasks, we characterize six streaming temporal formulations: *Cross-session understanding* and *Real-time Perception* each follow a unified setup, while *Adaptive Response* tasks adopt four distinct patterns (with TPG and MPA sharing one). **Notations:** ● denotes user instructions, such as forward questions or reminder registrations; TPG and MPA trigger autonomously without explicit user prompts. ● denotes probes used to evaluate whether the model should remain `SILENT` or `INTERVENE`. *Real-time Perception*: evidence and query co-occur in the current session; *Cross-session Understanding*: evidence comes from prior sessions; and *Adaptive Response*: evidence comes from the current session for intra-session tasks (ERA and RCR), or from prior sessions for inter-session tasks (MPA, PRM, and TPG).

features or tokens before passing them to the model (Wu et al., 2026; Liu et al., 2026a) or incrementally update stored representations as new frames arrive (Liu et al., 2026b; Qu et al., 2026). Structured memories organize history around events (Zeng et al., 2025; Liang et al., 2026) or objects and their state changes (Dong et al., 2026). KV-cache approaches retrieve or compress past states (Di et al., 2025; Chen et al., 2026b) and organize them hierarchically (Zhang et al., 2026a). Text-based approaches preserve reasoning traces (Wang et al., 2026; Liu et al., 2026c) or structured summaries (Jiang et al., 2026) for later use. Parametric memory systems update model parameters during streaming (Sun et al., 2026; Chen et al., 2026a). Yet most approaches manage context within a single continuous video. In real-world use, models need memory that can persist through interruptions, and remain useful when interactions resume. EgoMemo (Sitong et al., 2026), GROVE (Gong et al., 2026), and StreamMind (Zhang et al., 2026b) construct persistent memory online, but retrieval and reasoning over that memory introduce substantial response latency, weakening time-sensitive adaptive responses. Questions remain about when to use memory, whether persistent state can be stored affordably, and how to respond when required evidence is unavailable.

3 APM-BENCH

APM-Bench organizes egocentric video streams into multi-session trajectories of related activities, preserving continuous temporal order within sessions and realistic time gaps between them. The benchmark systematically evaluates three complementary capabilities: (1) *Cross-session Understanding* assesses whether persistent memory reliably retains essential information across completed sessions; (2) *Real-time Perception* evaluates streaming perception in the current scene, while examining whether incorporating persistent memory impacts real-time perception performance; and (3) *Adaptive Response* tests whether the model can bridge current scenes with persistent memory to deliver timely, proactive responses. Across these capabilities, we introduce 12 task types categorized

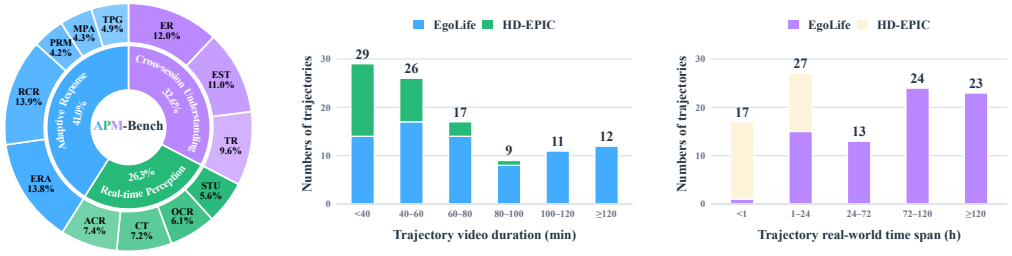

Figure 4: **APM-Bench statistics.** Left: distribution of 2,719 candidates across 12 tasks and three capability families. Middle: cumulative video duration across sessions in each trajectory. Right: wall-clock span from the first session start to the last session end, including inter-session gaps.

into intra-session or inter-session settings based on evidence location, where tasks in the first two families are formulated as multiple-choice questions, while all tasks in Adaptive Response are structured as open-ended questions. Furthermore, a dedicated evaluation set is constructed to examine whether models recognize when required evidence is unavailable rather than fabricate an answer.

3.1 Benchmark Construction

Data Source. APM-Bench builds on two egocentric datasets: EgoLife (Yang et al., 2025), with multi-day recordings, transcripts, and timestamped captions, and HD-EPIC (Perrett et al., 2025), with fine-grained action and object annotations for structured kitchen procedures.

Construction Pipeline. As illustrated in Figure 5, APM-Bench is constructed in two stages. In *Stage 1*, we organize EgoLife and HD-EPIC videos into activity-related multi-session trajectories with real-world timestamps. In *Stage 2*, we generate candidates for the three capability families and filter and refine them through automated and human review, yielding 2,719 candidates. On 300 sampled questions, two annotators reach a Cohen's kappa of 0.868 (Cohen, 1960).

Evidence Availability-Aware Evaluation Set. Finite storage and compression prevent persistent memory from retaining all visual history. We therefore curate 260 Cross-session Understanding questions to test whether models recognize unavailable evidence. Models access only the two most recent completed sessions before the query: 130 questions have all required evidence within this history, while the other 130 require evidence outside it. Each question includes a coarse time span and a fifth option indicating insufficient available evidence. This setting tests whether models answer when evidence is accessible and acknowledge when it is not.

3.2 Detail of APM-Bench

As illustrated in Figure 3, we characterize six streaming formulations of queries, evidence, spanning intra-session and inter-session settings based on evidence location.

Cross-session Understanding requires evidence from previous sessions. *Episodic Recall* (ER) recalls or summarizes events and activities from earlier sessions. *Entity State Tracking* (EST) tracks the states or locations of objects and other entities over time. *Temporal Reasoning* (TR) compares multiple historical moments to infer event orderings and temporal changes. Tasks in this family are formulated as multiple-choice questions and span single- and multi-evidence temporal grounding.

Real-time Perception evaluates understanding of the current visual scene within the ongoing session. *Action Recognition* (ACR) identifies actions performed by the wearer or nearby individuals. *Counting* (CT) counts instances of specified entities. *Optical Character Recognition* (OCR) reads visible text, labels, or screen content. *Spatial Understanding* (STU) determines spatial relationships among specific entities. Tasks in this family are also formulated as multiple-choice questions.

Adaptive Response evaluates whether the model delivers timely assistance when trigger conditions are met and remains silent otherwise. *Evidence-Ready Answering* (ERA) releases a multiple-choice question before its causal evidence appears; the model must remain `SILENT` until the evidence is sufficient, then `INTERVENE` with the selected option and rationale. *Registered-Condition Response*

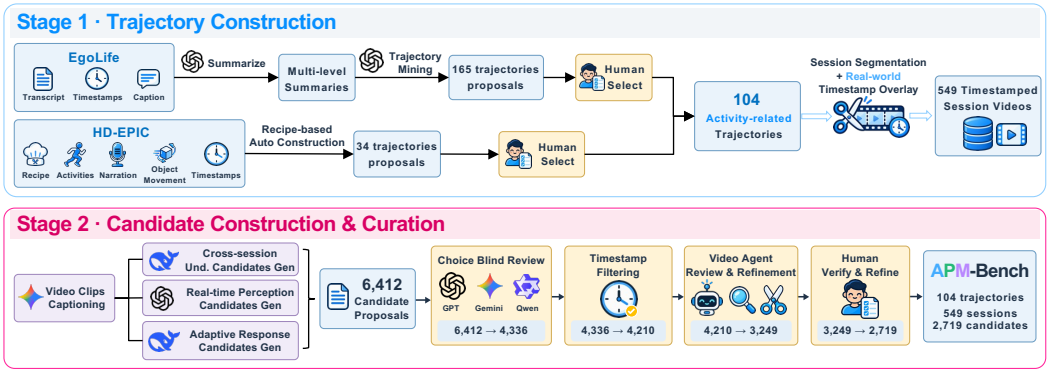

Figure 5: **APM-Bench construction pipeline.** In *Stage 1*, EgoLife and HD-EPIC are organized into activity-related trajectories and segmented into sessions; real-world timestamps from the source metadata are rendered onto session videos that lack visible timestamps. In *Stage 2*, candidates are generated from the session videos and progressively filtered and refined through choice-blind review, timestamp filtering, video-agent review, and human verification, yielding 2,719 candidates across 104 trajectories and 549 sessions. See Appendix D.1 for details.

(RCR) requires the model to detect whether the ongoing scene fulfills a reminder condition registered earlier. *Memory-Grounded Proactive Assistance* (MPA) leverages past experience without explicit user instructions to offer proactive guidance during related activities, requiring models to connect historical memory with current actions. *Proactive Reminder* (PRM) triggers a reminder registered in a prior session when conditions arise in a later session. *Task Progress Guidance* (TPG) tracks progress across long-running tasks and delivers task-relevant assistance upon resumption. ERA and RCR are *intra-session* tasks confined to the current session, whereas MPA, PRM, and TPG are *inter-session* tasks that bridge current visual events with persistent historical memory. Except for ERA, the other tasks require outputting the decision, rationale, and response. All adaptive tasks are evaluated in an open-ended format via LLM-as-a-judge.

As shown in Figure 4, each trajectory contains 69 minutes of video on average, with individual sessions averaging 13 minutes, while its real-world span can extend across hours or days due to inter-session gaps. This separation between video duration and elapsed real-world time reflects the intermittent interactions that persistent memory must support, with more statistics in Appendix C.

3.3 Evaluation Protocol and Metrics

Online Inference Protocol During inference, models access prior-session persistent memory and the current session's causal video prefix ending at the probe or query timestamp, strictly adhering to causal constraints. For Adaptive Response, each candidate contains probes at precise timestamps labeled as `SILENT` or `INTERVENE`, treating each probe as an independent runtime instance (3,768 in total). Specifically, ERA releases its multiple-choice question at session start without repeating it at subsequent probes; for MPA and TPG, probes provide brief task definitions to guide decisions; RCR and PRM inject textual reminder instructions at registration timestamps. Conversely, candidates in Cross-session Understanding (887) and Real-time Perception (716) each form a single runtime instance evaluated at the query timestamp, requiring the model to directly output option choices.

Metrics Table 2 summarizes the evaluation metrics. For *Adaptive Response*, we propose a Gated LLM-Judge Score across all five tasks. A response passes the gate if it correctly decides to `INTERVENE` or remain `SILENT`; for positive ERA probes, the predicted MCQA option must additionally match the ground truth. Probes failing the gate receive $g_r = 0$. For gate-passing responses, DeepSeek-V4-Flash evaluates the rationale against the reference response and assigns a score $g_r \in \{1, \ldots, 5\}$.For candidate c, let $\mathcal{P}_c$ and $\mathcal{N}_c$ denote its positive and negative probe sets, with mean scores $\bar{g}_c^+ = \frac{1}{|\mathcal{P}_c|} \sum_{r \in \mathcal{P}_c} g_r$ and $\bar{g}_c^- = \frac{1}{|\mathcal{N}_c|} \sum_{r \in \mathcal{N}_c} g_r$, respectively. The candidate-level score s_c and the task-level score over the candidate set C_t are defined as:

Table 2: Evaluation metrics across capability families and the *Evidence Availability-Aware* setting.

Evaluation	Output Format	Metric
Cross-session Understanding	4-way MCQA	Accuracy
Real-time Perception	4-way MCQA	Accuracy
Evidence Availability-Aware	5-way MCQA	Accuracy
Adaptive Response	Open-ended	Gated LLM-Judge

$$\mathrm{GatedLLMJudge}_t = \frac{20}{|C_t|}\sum_{c\in C_t} s_c, \quad s_c = \begin{cases} \frac{1}{2}(\bar{g}_c^{+} + \bar{g}_c^{-}), & \mathcal{P}_c \neq \emptyset \wedge \mathcal{N}_c \neq \emptyset, \\ \bar{g}_c^{\star}, & \text{otherwise.} \end{cases} \tag{1}$$

where $\bar{g}_c^{\star}$ denotes the mean score over the sole available probe set, either $\mathcal{P}_c$ or $\mathcal{N}_c$. The factor of 20 converts candidate scores to a 0–100 task score. See Appendix A.3 for the judging rubric.

Storage cost quantifies the persistent state retained across completed prior sessions, strictly excluding the ongoing session. For each trajectory τ, we capture the peak prior-session storage footprint B_τ^{peak} normalized by the cumulative prior-session duration D_τ^{peak} in hours:

$$\text{Storage/hr} = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \frac{B_\tau^{\text{peak}}}{D_\tau^{\text{peak}}}. \tag{2}$$

4 Experiment

4.1 Experiment Setup

General Large Video Models We evaluate a no-memory baseline and two memory conditions for general video models. *SimpleStream* uses Qwen3-VL-8B-Instruct (Bai et al., 2025) with only the four most recent frames as the no-memory baseline. *Raw Video as Memory* stores all prior-session videos and replays them at query time. *Text Summary as Memory* generates a summary at the end of each session and provides prior-session summaries together with the current causal video prefix at query time. All models use 1 FPS video input. Proprietary models and Qwen-series models uniformly sample up to 1,024 frames, while InternVL3.5-8B (Wang et al., 2025) and VideoLLaMA3-7B (Zhang et al., 2025a) sample up to 128 frames.

Specialized Streaming Memory Systems We evaluate eight methods across five memory representations: *KV Cache*, *Visual Tokens/Features*, *Event Tree*, *Parametric Memory*, and *Reasoning Thoughts*. All use official settings and process prior sessions sequentially. For methods supporting streaming input, latency is measured from the last required memory update to the first output token; otherwise, from processing the causal video prefix to the first output token. As most methods target single continuous videos without persistent-state export, Table 3 estimates storage from the memory state maintained during inference. Implementation details are in Appendix A.1 and A.2.

4.2 Main Results

What to store matters: ***raw video performs best, while event-structured methods lead among specialized memory systems.*** Tables 3 and 4 show that richer retained information benefits Cross-session Understanding. Without prior-session memory, *SimpleStream* performs substantially worse on cross-session tasks. *Raw Video as Memory* preserves the richest visual evidence and achieves the strongest cross-session performance, whereas *Text Summary as Memory* loses visual details during summarization and degrades cross-session understanding. Specialized systems use different representations, including KV caches, visual tokens or features, event structures, parametric memory, and reasoning thoughts. Among them, ***event-structured*** methods achieve the strongest utility, suggesting that organizing history around events is a promising alternative to raw-video replay. ***Parametric memory*** is also appealing because its state does not grow with video length, although the evaluated method still suffers from limited utility and nontrivial latency. By organizing intermittent interactions into sessions, APM-Bench naturally defines memory-storage boundaries and enables clearer comparison across memory representations, as further analyzed in Appendix B.6.

Table 3: Task performance and efficiency of the evaluated memory methods. Streaming and Persistent indicate support for streaming input and persistent memory, respectively. CS Und.: *Cross-session Understanding*; Perception: *Real-time Perception*; Adaptive: *Adaptive Response*. Storage cost is reported per video hour, and Overall is the mean of the three capability scores.

Method	Streaming	Persistent	CS Und.	Perception	Adaptive	TTFT (s)	Storage Cost	Overall
w/o Memory								
SimpleStream (Recent-4)	✗	✗	27.21	53.06	32.48	0.56	–	37.58
Raw Video as Memory								
Seed-2.0-Lite (ByteDance Seed, 2026)	–	–	54.47	50.75	41.88	–		49.03
Gemini 3.6 Flash (Google DeepMind, 2026)	–	–	**69.37**	64.20	47.05	–	3.01 GiB	**60.21**
Qwen3.8-27B (Qwen Team, 2026)	–	–	50.10	48.84	38.31	30.72		45.75
Text Summary as Memory								
Seed-2.0-Lite (ByteDance Seed, 2026)	–	–	45.34	52.44	44.08	–	7.75 KiB	47.29
Gemini 3.6 Flash (Google DeepMind, 2026)	–	–	47.50	**65.98**	**47.14**	–	3.09 KiB	53.54
Qwen3.8-27B (Qwen Team, 2026)	–	–	41.05	48.19	37.90	8.13	6.84 KiB	42.38
KV Cache								
HERMES (Zhang et al., 2026a)	✓	✗	34.12	42.04	23.05	4.35	0.67 GiB	33.07
ReKV (Di et al., 2025)	✓	✗	31.24	32.90	18.80	2.34	20.84 GiB	27.65
Visual Tokens / Features								
FLUXMem (Xie et al., 2026)	✗	✗	35.98	27.59	21.63	22.85	0.33 GiB	28.40
Flash-VStream (Zhang et al., 2025b)	✗	✗	31.89	30.76	11.41	8.17	1.13 GiB	24.69
Event Tree								
StreamForest (Zeng et al., 2025)	✗	✗	**37.79**	43.62	6.48	36.15	5.34 GiB	29.30
OASIS (Liang et al., 2026)	✓	✓	35.46	**52.92**	**35.99**	32.44	2.66 GiB	**41.46**
Parametric Memory								
Video-Salmon-S (Sun et al., 2026)	✗	✗	15.17	22.87	18.12	9.49	1.81 GiB	18.72
Reasoning Thoughts								
VST (Guan et al., 2026)	✓	✗	20.63	16.75	15.24	**1.82**	**11.64 KiB**	17.54

Efficiency analysis: ***most specialized memory systems struggle to balance utility, latency, and storage.*** The activity-related trajectory design preserves long real-world spans while reducing the amount of video history, making storage and latency easier to measure across methods. Table 3 shows that storage and latency are closely coupled with utility. *Raw-video memory* achieves strong performance but requires GiB-scale storage and incurs high query-time overhead, while *text summaries* reduce storage to the KiB scale at the cost of cross-session performance. Specialized memory systems improve efficiency in different ways, but compact or fast methods often sacrifice utility, while stronger systems can require substantially more storage or response time. These results reveal a clear utility–latency–storage trade-off, highlighting the need to ***jointly consider all three dimensions when designing persistent memory for real-world streaming assistants***.

When to use memory: ***not every interaction requires persistent memory, especially real-time perception.*** Real-time Perception depends on the current visual scene and often does not require prior-session history. As shown in Table 4, replacing raw-video history with text summaries improves Real-time Perception for most general video models, suggesting that excessive or irrelevant history can interfere with current-scene understanding. This reveals a tension between maintaining rich historical context and accurate real-time perception: information useful for later recall may be unnecessary or distracting in the current interaction.Retaining more history is therefore insufficient; the system must also determine whether that history is useful now and what should be exposed to the model. Persistent memory should therefore be used selectively, ***injecting historical information only when it is relevant to the ongoing interaction***.

Evidence availability awareness: ***answering and recognizing missing evidence are distinct capabilities.*** On the *Evidence Availability-Aware* evaluation, Gemini 3.6 Flash performs strongly when the required evidence remains accessible but is less reliable at recognizing when it is unavailable. SimpleStream shows the opposite pattern: with only the four most recent frames, it strongly favors the insufficient-evidence option but rarely answers correctly when historical evidence is available. These contrasting cases show that successful recall does not imply reliable awareness of what evidence remains accessible. One practical direction is to ***make the coverage of retained history explicit***, enabling the assistant to recognize when required evidence is unavailable rather than fabricate an answer. Further analysis is provided in Appendix B.7.

Table 4: Task-level results across the three capability families and *Evidence Availability-Aware evaluation*. EA: evidence available in accessible session. EU: evidence unavailable detection.

Model / Method	Cross-session Und.			Real-time Perception				Adaptive Response					Overall	Evidence Availability-Aware	
	ER	EST	TR	ACR	CT	OCR	STU	ERA	RCR	MPA	PRM	TPG		EA	EU
w/o Memory															
SimpleStream	27.52	28.52	25.57	56.44	35.90	73.49	46.41	50.89	30.22	26.54	32.11	22.64	37.58	3.85	95.38
Raw Video as Memory															
Seed-2.0-Lite	54.13	63.09	46.18	57.43	43.59	53.61	48.37	41.97	65.63	24.45	51.56	25.80	49.03	64.62	66.15
Gemini 3.6 Flash	**73.70**	**71.81**	**62.60**	67.33	42.05	**87.95**	**59.48**	47.73	73.62	26.52	**61.80**	25.58	**60.21**	**85.38**	59.23
Qwen3.8-27B	54.74	48.99	46.56	55.94	34.87	62.05	42.48	44.38	53.33	26.81	41.75	25.27	45.75	59.23	46.92
Qwen3-VL-8B-Instruct	40.98	46.64	26.72	45.54	36.41	62.65	37.91	37.53	29.83	25.89	29.63	25.01	37.77	40.00	**66.92**
InternVL3.5-8B	32.42	43.96	27.86	41.58	35.90	43.98	36.60	23.06	24.46	14.83	16.84	18.24	31.25	50.77	40.77
VideoLLaMA3-7B	22.02	27.18	25.95	28.22	24.10	21.08	32.68	11.98	25.36	14.45	16.32	17.64	22.91	20.00	49.23
Text Summary as Memory															
Seed-2.0-Lite	49.54	48.32	38.17	62.38	43.59	55.42	48.37	49.39	71.15	25.49	46.18	**28.19**	47.29	–	–
Gemini 3.6 Flash	49.85	50.67	41.98	**71.29**	**48.21**	84.94	**59.48**	**54.20**	**75.68**	**28.82**	51.01	26.01	53.54	–	–
Qwen3.8-27B	45.26	38.59	39.31	52.97	34.36	59.04	46.41	41.59	56.63	27.18	38.11	25.98	42.38	–	–
Qwen3-VL-8B-Instruct	30.58	39.60	24.81	48.02	33.85	59.64	42.48	38.35	34.25	23.32	31.01	25.64	36.06	–	–
InternVL3.5-8B	25.69	31.21	27.48	45.54	37.95	54.22	39.87	27.34	29.93	15.00	17.44	18.80	31.41	–	–
VideoLLaMA3-7B	19.88	21.14	28.24	35.64	31.28	25.30	32.03	11.66	28.35	15.86	13.03	17.11	23.78	–	–
KV Cache															
HERMES	36.09	34.23	32.06	39.60	33.85	69.88	24.84	35.82	24.64	17.39	19.08	18.34	33.07	7.69	**90.77**
ReKV	27.52	34.90	31.30	30.69	34.36	39.76	26.80	22.72	15.26	17.00	19.75	19.28	27.65	10.77	63.85
Visual Tokens / Features															
FLUXMem	32.42	41.95	33.59	26.73	20.51	34.34	28.76	29.25	23.13	17.17	20.04	18.55	28.40	15.38	71.54
Flash-VStream	29.97	38.59	27.10	31.19	33.85	32.53	25.49	5.32	17.70	10.82	13.07	10.15	24.69	7.69	52.31
Event Tree															
StreamForest	34.86	**42.62**	**35.88**	45.05	**42.05**	56.02	31.37	8.42	10.79	4.01	6.90	2.30	29.30	15.38	45.38
OASIS	**36.39**	40.60	29.39	**57.43**	36.41	**74.70**	**43.14**	**52.93**	**52.98**	**22.07**	**28.90**	**23.06**	**41.46**	**23.08**	86.15
Parametric Memory															
Video-Salmon-S	17.43	15.10	12.98	17.33	20.00	31.93	22.22	22.61	23.22	15.16	14.65	14.98	18.72	16.15	25.38
Reasoning Thoughts															
VST	20.80	23.15	17.94	10.40	29.23	16.27	11.11	5.00	17.65	16.62	20.48	16.42	17.54	10.77	55.38

4.3 Why Adaptive Response Remains Difficult

Adaptive Response remains the most challenging capability even for the strongest general video models. The task-level results in Table 4 show a clear distinction between assistance driven by explicit registrations and fully autonomous proactive assistance. RCR and PRM provide the model with a previously registered condition or reminder, giving it a concrete target to monitor in the current scene. Models perform substantially better under these paradigms than on MPA and TPG, where no explicit trigger specifies which historical experience should become relevant.

MPA and TPG require models to connect relevant past experience to the current scene, decide whether to intervene, and provide useful assistance. Richer history alone does not solve this: even with Raw Video as Memory, the strongest general model remains weak on these tasks despite much stronger Cross-session Understanding. Further analysis is provided in Appendix B.1 and B.2.

5 Conclusion

We introduce APM-Bench, which organizes egocentric experience into activity-related multi-session trajectories to simulate intermittent real-world use and evaluate persistent memory in streaming video models. A useful persistent memory must be storable and reusable across sessions, while carefully balancing what to store, when to use it, and efficiency. Our experiments reveal clear trade-offs. Raw video preserves the richest visual evidence and achieves the strongest cross-session performance, but incurs substantial storage and latency costs. Compact representations reduce these costs but may lose details needed later; among specialized systems, event-structured memory shows the strongest utility. Excessive history can impair real-time perception, motivating selective memory use. Moreover, strong recall ability does not guarantee awareness when required evidence is unavailable, and adaptive response remains challenging even with rich historical context. Overall, current methods still struggle to simultaneously achieve reliable long-term recall, selective memory use, low storage and latency overhead, and effective proactive assistance across sessions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with dataset construction and generation, code debugging, manuscript editing for readability and clarity, and the preparation of figures and experimental scripts. All AI-assisted outputs were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, data, code, and artifacts produced with the aid of generative AI.

REFERENCES

Ant Group. Realtime-Venus: A full-duplex interaction system with asynchronous delegation, 2026. URL https://arxiv.org/abs/2609.13814v3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report, 2025. URL https://arxiv.org/abs/2511.21631.

ByteDance Seed. Seed2.0 Model Card: Towards Intelligence Frontier for Real-World Complexity, 2026. URL https://arxiv.org/abs/2607.00248.

Joya Chen, Zeyun Zhong, and Mike Zheng Shou. StreamTTT: Reconciling Real-Time Perception and Long-Term Memory in Streaming VLMs, 2026a. URL https://arxiv.org/abs/2608.13416.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. StreamingTOM: Streaming Token Compression for Efficient Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24675–24685, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_StreamingTOM_Streaming_Token_Compression_for_Efficient_Video_Understanding_CVPR_2026_paper.html.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/67a9b444cbcd647572c88194619f72d5-Abstract-Conference.html.

Mingkang Dong, Muxin Pu, Jie Li, Bohan Guo, Songruo Chen, Bin Ren, Xu Zheng, Chen Zhao, Tianwen Qian, Mohamed Elhoseiny, and Yuqian Fu. ObjectStream: Latent Objects as Memory Anchors for Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2607.28312.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. EGOSTREAM: A Diagnostic Benchmark for Streaming Episodic Memory in Egocentric Vision, 2026. URL https://arxiv.org/abs/2605.31557.

Haonan Ge, Yiwei Wang, Hang Wu, and Yujun Cai. What Should a Streaming Video Model Remember?, 2026. URL https://arxiv.org/abs/2606.16353.

Sitong Gong, Caixin Kang, Tianyu Yan, Guo Chen, Bo Zheng, Kaipeng Zhang, Yunzhi Zhuge, Xiang Ruan, Huchuan Lu, and Yifei Huang. GROVE: Growing and Reasoning over Temporally Stratified Memory from Streaming Video Experience, 2026. URL https://arxiv.org/abs/2608.02392.

Google DeepMind. Gemini 3.6 Flash Model Card. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/.

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, 2026. URL https://arxiv.org/abs/2603.12262.

Xinru Jiang, Lin Zhao, Xi Xiao, Yunbei Zhang, Janet Wang, Chenrui Ma, Haolin Li, Yanzhi Wang, Yifan Gong, and Octavia Camps. Dynamic Hub-and-Spoke Memory for Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2608.30294.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. EgoSAT: A Comprehensive Benchmark of Egocentric Streaming Interaction Understanding, 2026. URL https://arxiv.org/abs/2606.24422.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding?, 2025. URL https://arxiv.org/abs/2501.05510.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. OASIS: On-Demand Hierarchical Event Memory for Streaming Video Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2821–2831, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liang_OASIS_On-Demand_Hierarchical_Event_Memory_for_Streaming_Video_Reasoning_CVPR_2026_paper.html.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding, 2024. URL https://arxiv.org/abs/2411.03628.

Jinming Liu, Jianguo Huang, Zhaoyang Jia, Jiahao Li, Xiaoyi Zhang, Zongyu Guo, Bin Li, Wenjun Zeng, Yan Lu, and Xin Jin. An Efficient Streaming Video Understanding Framework with Agentic Control, 2026a. URL https://arxiv.org/abs/2605.17921.

Yuxin Liu, Peiqin Zhuang, and Yali Wang. StreamEMS: Streaming Video Understanding with Self-Evolving Memory Scheme for Vision-Language Models, 2026b. URL https://arxiv.org/abs/2608.27881.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in Streaming Video, 2026c. URL https://arxiv.org/abs/2603.12938.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. PhoStream: Benchmarking Real-World Streaming for Omnimodal Assistants in Mobile Scenarios, 2026. URL https://arxiv.org/abs/2601.22575.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, 2025. URL https://arxiv.org/abs/2502.04144.

Hongyu Qu, Guangming Yao, Ling Xing, Xiaobin Hu, Rongxing Ding, Guibin Zhang, Fan Zhang, Yi Yuan, Xiangbo Shu, and Shuicheng Yan. Beyond Retrieval: Progressive Latent Memory Evolution for Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2609.04131.

Qwen Team. Qwen3.8-27B Model Card. Model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Dongchuan Ran, Linyu Ou, Xueheng Li, Wenwen Tong, Chenxu Guo, Hewei Guo, Kaibing Wang, and Lewei Lu. EgoPro-Bench: Benchmarking Personalized Proactive Interaction in Egocentric Video Streams, 2026. URL https://arxiv.org/abs/2605.07299.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A Simple Baseline for Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2604.02317.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. RIVER: A Real-Time Interaction Benchmark for Video LLMs, 2026. URL https://arxiv.org/abs/2603.03985.

Gong Sitong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing Proactive Assistance in Continuous Egocentric Videos, 2026. URL https://arxiv.org/abs/2607.11523.

Guangzhi Sun, Yixuan Li, Xiaodong Wu, Yudong Yang, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN S: Memory-Enhanced Streaming Audio-Visual LLM, 2026. URL https://arxiv.org/abs/2510.11129.

Lu Wang, Zhuoran Jin, Yupu Hao, Yubo Chen, Kang Liu, Yulong Ao, and Jun Zhao. Think While Watching: Online Streaming Segment-Level Memory for Multi-Turn Video Reasoning in Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2603.11896.

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, 2025. URL https://arxiv.org/abs/2508.18265.

Hang Wu, Sherin Mary Mathews, Yujun Cai, Ming-Hsuan Yang, and Yiwei Wang. Semantic-Aware Adaptive Visual Memory for Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2605.07897.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. FluxMem: Adaptive Hierarchical Memory for Streaming Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Xie_FluxMem_Adaptive_Hierarchical_Memory_for_Streaming_Video_Understanding_CVPR_2026_paper.pdf.

ShuHang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, LingHao Zhang, Shikang Wang, Yixin Liu, Hanbo Zhang, Ying Ma, and Xuming Hu. RTV-Bench: Benchmarking MLLM Continuous Perception, Understanding and Reasoning through Real-Time Video. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0600. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19e4ea30dded58259665db375885e412-Abstract-Datasets_and_Benchmarks_Track.html.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-Time Vision-Language Interaction Intelligence, 2026. URL https://arxiv.org/abs/2606.14777.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient Online Video Understanding with Persistent Event Memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6dd91fec726dbed8915a1fbadd91d1d2-Abstract-Conference.html.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, 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, 2025a. URL https://arxiv.org/abs/2501.13106.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

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, 2026a. doi: 10.18653/v1/2026.acl-long.381. URL https://aclanthology.org/2026.acl-long.381/.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. StreamArena: Toward Continuous, Interactive, and Long-Horizon Agentic Streaming Video Understanding, 2026b. URL https://arxiv.org/abs/2608.05703.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. Eyes Wide Open: Ego Proactive Video-LLM for Streaming Video, 2025c. URL https://arxiv.org/abs/2510.14560.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing LYU, and Xirong Li. OmniPro: A Comprehensive Benchmark for Omni-Proactive Streaming Video Understanding, 2026. URL https://arxiv.org/abs/2605.18577.

APPENDIX CONTENTS

A Implementation Details . . . 15

A.1 Compute . . . 15

A.2 Backbones . . . 15

A.3 LLM-as-Judge rubric . . . 15

B Additional Results and Analysis . . . 16

B.1 Adaptive decision accuracy . . . 16

B.2 Response quality after correct decisions . . . 17

B.3 Latency . . . 19

B.4 Persistent storage . . . 19

B.5 Instruction following . . . 20

B.6 Effect of trajectory organization . . . 20

B.7 Evidence availability and answer accuracy . . . 21

B.8 Task-level performance . . . 22

C Dataset Statistics . . . 24

C.1 Trajectory activities . . . 24

C.2 Evidence composition . . . 24

C.3 Evidence distance measured in sessions . . . 24

C.4 Real-world span and organized trajectory duration . . . 25

D Data Construction and Annotation Details . . . 25

D.1 Dataset construction . . . 25

D.2 Human verification . . . 26

D.3 MCQA annotation format . . . 26

D.4 ERA: evidence readiness . . . 26

D.5 RCR: in-session conditional reminder . . . 28

D.6 MPA: assistance from prior experience . . . 28

D.7 PRM: cross-session registered reminder . . . 28

D.8 TPG: continuing-task guidance . . . 29

E Task Prompts . . . 29

E.1 System prompt . . . 29

E.2 MCQA prompts . . . 29

E.3 Adaptive Response probes . . . 30

E.4 Memory Loss-Aware prompt . . . 32

E.5 Text-summary memory prompt . . . 33

A Implementation Details

A.1 Compute

All experiments ran on H20 GPUs with the same host configuration, enabling a fair comparison of latency and storage cost across open-source video models and specialized memory systems.

A.2 Backbones

Table 5 lists the backbone of each specialized system and SimpleStream.

Table 5: Visual-language backbones of specialized memory systems and SimpleStream.

System	Visual-language backbone
SimpleStream	Qwen3-VL-8B-Instruct
HERMES	Qwen2.5-VL-7B-Instruct
ReKV	LLaVA-OneVision-Qwen2-7B
FluxMem	Qwen2.5-VL-7B-Instruct
Flash-VStream	Qwen2-VL-7B
StreamForest	Qwen2-7B
OASIS	Qwen3-VL-8B-Instruct
Video-SALMONN S	Qwen3-VL-8B
VST	Qwen2.5-VL-7B-Instruct

A.3 LLM-as-Judge rubric

After the deterministic gate in Section 3.3, DeepSeek V4 Flash scores response quality using the rubric and task guidance below. The judge receives the causal cutoff, reference annotations, and model response for each probe.

DeepSeek V4 Flash: complete judge system rubric and output contract

```
You are an impartial evaluator for a streaming video assistant
    benchmark.

The assistant's SILENT/INTERVENE decision has already passed
    deterministic correctness checks. Evaluate only the semantic
    quality of its explanation and, when it intervenes, its proactive
    response. Reference annotations describe accepted evidence and
    one acceptable response; do not require lexical overlap or
    identical wording. Do not reward verbosity. Do not infer facts
    outside the supplied annotations.

For a correct SILENT decision, judge whether the explanation
    identifies the decisive condition that is still missing and
    distinguishes this evaluation window from a valid response
    opportunity. For ERA, evaluate only whether the reason correctly
    explains whether the answer evidence is causally available; the
    MCQA answer itself has already passed deterministic checking.

Use this 1-5 scale:
5 = task-faithful, fully grounded, factually accurate, precise, and
    useful.
4 = core content is correct and useful, with only a minor omission or
    imprecision.
3 = basically correct but generic, incomplete, or uses only part of
    the important evidence.
2 = related but has a major grounding gap, factual issue, or omission
    that could mislead.
```

```
1 = the decision happens to be correct, but the explanation or
    response is substantially inconsistent or unusable.

Return exactly one JSON object with all and only the following fields:
{
  "score": 1,
  "criterion_checks": {
    "task_fidelity": "pass | partial | fail",
    "historical_or_registration_grounding": "pass | partial | fail |
    not_applicable",
    "current_trigger_or_silence_grounding": "pass | partial | fail",
    "factual_support": "pass | partial | fail",
    "usefulness": "pass | partial | fail | not_applicable"
  },
  "critical_issues": [],
  "justification": "Concise evidence-based explanation."
}

The value shown as 1 for score is an example; replace it with one
    integer from 1 through 5. For each criterion, return exactly one
    of the literal enum values separated by | above, not the whole
    displayed string. Include every criterion_checks key even when
    its value is not_applicable. Keep critical_issues short and use
    an empty array when there is no critical issue. Keep
    justification concise and evidence-based. Do not wrap the JSON in
    Markdown or add text before or after it.
```

DeepSeek V4 Flash: task-specific guidance

```
Task-specific guidance field (select the matching task):

ERA: Check whether the reason correctly explains that answer evidence
    is or is not causally available at this heartbeat.

RCR: Check fidelity to the in-session registration and whether the
    current observable condition justifies the registered response.

MPA: Check whether earlier experience materially improves assistance
    for the current scene and whether the response is actionable
    rather than generic.

PRM: Check fidelity to the earlier registration, whether its
    observable condition is satisfied, and whether the response
    preserves the obligation.

TPG: Check whether history and the current scene belong to the same
    continuing task and whether progress or next-step guidance is
    accurate.
```

B Additional Results and Analysis

B.1 Adaptive decision accuracy

For task t, let P_t be the fraction of positive probes with a correct `INTERVENE` decision and N_t the fraction of negative probes with a correct `SILENT` decision. Decision balanced accuracy is $(P_t + N_t)/2$. The overall rate pools probes across the five Adaptive tasks before averaging the positive and negative rates. This measure tests whether the model acts at the right time; the main Gated Judge score also evaluates the answer or response under the Section 3.3 protocol. ERA's

positive decision rate alone does not require the MCQA option to be correct. Table 6 reports the results.

Table 6: Adaptive decision accuracy (%). Each task score averages the correct `INTERVENE` rate on positive probes and the correct `SILENT` rate on negative probes. Overall pools probes across the five tasks before reporting the positive rate (P), negative rate (N), and their balanced average (BA).

Model / Method	ERA	RCR	MPA	PRM	TPG	Overall $P/N/BA$
w/o Memory						
SimpleStream	65.89	50.81	54.53	50.86	50.39	19.86/92.81/56.34
Raw Video as Memory						
Seed-2.0-Lite	62.10	73.63	53.94	69.56	58.15	80.66/46.65/63.66
Gemini 3.6 Flash	63.80	82.76	51.58	77.55	52.16	74.33/59.04/66.69
Qwen3-VL-8B	58.49	51.47	56.80	52.58	58.18	64.79/44.28/54.54
Qwen3.8-27B	68.44	67.62	52.37	59.97	52.62	53.69/71.01/62.35
InternVL3.5-8B	52.42	50.00	50.75	50.86	52.77	90.37/13.42/51.90
VideoLLaMA3-7B	37.62	51.10	51.69	50.31	51.12	84.65/9.33/46.99
Text Summary as Memory						
Seed-2.0-Lite	67.77	78.30	53.29	65.84	54.65	77.10/55.60/66.35
Gemini 3.6 Flash	70.39	84.21	54.56	70.53	51.56	71.55/66.88/69.22
Qwen3-VL-8B	60.27	53.41	58.21	57.15	60.53	74.33/41.26/57.79
Qwen3.8-27B	68.07	67.53	55.33	62.52	53.40	64.53/61.26/62.89
InternVL3.5-8B	53.19	49.69	49.80	51.49	50.94	80.66/24.51/52.59
VideoLLaMA3-7B	40.71	50.00	50.28	51.40	50.88	87.51/7.72/47.62
KV Cache						
HERMES	50.88	54.26	53.49	52.75	52.34	58.02/51.97/55.00
ReKV	52.80	35.67	47.29	42.60	49.77	52.99/34.26/43.63
Visual Tokens / Features						
FluxMem	55.54	52.29	53.44	46.86	51.97	74.67/32.35/53.51
Flash-VStream	36.47	50.28	49.96	49.77	48.82	24.11/61.38/42.74
Event Tree						
StreamForest	40.46	21.30	16.27	22.62	11.32	23.33/27.72/25.53
OASIS	72.87	69.28	57.90	54.69	56.97	74.15/60.84/67.50
Parametric Memory						
Video-SALMONN S	53.22	53.96	50.75	46.07	52.64	68.78/34.34/51.56
Reasoning Thoughts						
VST	11.45	46.17	51.42	53.61	48.77	61.49/9.06/35.28

B.2 Response quality after correct decisions

After a probe passes the Section 3.3 gate, we measure the quality of the answer or proactive response when the model speaks, and its reason for remaining silent otherwise. For each candidate, we average judged probes within each polarity, weight the available polarities equally, and then average

over candidates with at least one judged probe. The judge's 1–5 score is converted to a 0–100 scale. Let A_c be the set of polarities with a judged, gate-passing probe for candidate c. Then

$$\text{ResponseQuality}_t = \frac{100}{5|C'_t|} \sum_{c \in C'_t} \frac{1}{|A_c|} \sum_{a \in A_c} \bar{j}_{c,a}, \tag{3}$$

where C'_t contains candidates with $A_c \neq \emptyset$ and $\bar{j}_{c,a}$ is the mean judge score for candidate c and polarity a. Table 7 reports these scores separately from the main Gated Judge result. Its superscripts give the balanced gate pass rate: we compute the fraction of positive and negative probes passing the deterministic gate separately, then average the two fractions. For ERA, a positive probe passes only if the model intervenes and selects the correct MCQA option.

Table 7: Response quality after the Section 3.3 gate (%). Judged probes are averaged within each available polarity, then by candidate, using Eq. 3. Superscripts give the balanced gate pass rate for each task: the mean of the positive and negative probe pass rates. Positive ERA probes also require the correct MCQA answer. Average is the mean of the five task scores.

Model / Method	ERA	RCR	MPA	PRM	TPG	Average
w/o Memory						
SimpleStream	$84.45^{59.1\%}$	$60.18^{50.8\%}$	$49.04^{54.5\%}$	$63.81^{50.9\%}$	$48.11^{50.4\%}$	61.12
Raw Video as Memory						
Seed-2.0-Lite	$92.91^{45.4\%}$	$88.52^{73.6\%}$	$46.45^{53.9\%}$	$72.52^{69.6\%}$	$44.73^{58.2\%}$	69.02
Gemini 3.6 Flash	$88.33^{54.2\%}$	$88.84^{82.8\%}$	$52.58^{51.6\%}$	$78.64^{77.6\%}$	$52.31^{52.2\%}$	72.14
Qwen3-VL-8B	$78.80^{47.6\%}$	$57.46^{51.5\%}$	$46.47^{56.8\%}$	$54.05^{52.6\%}$	$43.84^{58.2\%}$	56.12
Qwen3.8-27B	$83.84^{52.3\%}$	$77.35^{67.6\%}$	$52.23^{52.4\%}$	$66.78^{60.0\%}$	$50.64^{52.6\%}$	66.17
InternVL3.5-8B	$83.28^{27.9\%}$	$48.92^{50.0\%}$	$29.04^{50.7\%}$	$33.00^{50.9\%}$	$31.98^{52.8\%}$	45.24
VideoLLaMA3-7B	$72.50^{16.8\%}$	$50.13^{51.1\%}$	$27.74^{51.7\%}$	$32.01^{50.3\%}$	$32.12^{51.1\%}$	42.90
Text Summary as Memory						
Seed-2.0-Lite	$94.22^{52.6\%}$	$90.27^{78.3\%}$	$48.24^{53.3\%}$	$69.30^{65.8\%}$	$53.07^{54.7\%}$	71.02
Gemini 3.6 Flash	$86.59^{62.9\%}$	$89.51^{84.2\%}$	$52.85^{54.6\%}$	$71.70^{70.5\%}$	$54.83^{51.6\%}$	71.10
Qwen3-VL-8B	$79.77^{47.7\%}$	$63.83^{53.4\%}$	$41.88^{58.2\%}$	$53.41^{57.2\%}$	$40.74^{60.5\%}$	55.93
Qwen3.8-27B	$84.84^{48.9\%}$	$82.22^{67.5\%}$	$51.01^{55.3\%}$	$59.32^{62.5\%}$	$51.37^{53.4\%}$	65.75
InternVL3.5-8B	$78.86^{34.3\%}$	$60.41^{49.7\%}$	$29.83^{49.8\%}$	$33.58^{51.5\%}$	$33.51^{50.9\%}$	47.24
VideoLLaMA3-7B	$72.66^{16.3\%}$	$56.95^{50.0\%}$	$31.64^{50.3\%}$	$25.22^{51.4\%}$	$30.64^{50.9\%}$	43.42
KV Cache						
HERMES	$71.16^{49.9\%}$	$44.67^{54.3\%}$	$33.42^{53.5\%}$	$35.90^{52.8\%}$	$33.80^{52.3\%}$	43.79
ReKV	$68.51^{33.5\%}$	$42.82^{35.7\%}$	$37.31^{47.3\%}$	$47.95^{42.6\%}$	$40.56^{49.8\%}$	47.43
Visual Tokens / Features						
FluxMem	$72.50^{39.8\%}$	$44.14^{52.3\%}$	$31.61^{53.4\%}$	$43.42^{46.9\%}$	$34.53^{52.0\%}$	45.24
Flash-VStream	$42.08^{12.9\%}$	$34.88^{50.3\%}$	$21.48^{50.0\%}$	$26.23^{49.8\%}$	$22.38^{48.8\%}$	29.41
Event Tree						
StreamForest	$32.62^{25.3\%}$	$51.85^{21.3\%}$	$24.38^{16.3\%}$	$30.83^{22.6\%}$	$21.76^{11.3\%}$	32.29
OASIS	$85.22^{61.7\%}$	$76.11^{69.3\%}$	$39.53^{57.9\%}$	$51.39^{54.7\%}$	$40.77^{57.0\%}$	58.61
Parametric Memory						
Video-SALMONN S	$78.75^{28.8\%}$	$41.72^{54.0\%}$	$32.30^{50.7\%}$	$31.31^{46.1\%}$	$30.88^{52.6\%}$	42.99
Reasoning Thoughts						
VST	$72.24^{6.7\%}$	$38.23^{46.2\%}$	$32.72^{51.4\%}$	$38.17^{53.6\%}$	$32.20^{48.8\%}$	42.71

B.3 LATENCY

For local model runs, query-to-first-token time includes memory preparation, visual processing, and the time until the first generated token, under the Section 4.1 input settings. We additionally report an offline replay real-time factor, defined as the processing time from the beginning of a causal video prefix to the first output token divided by that prefix's video duration; video decoding is excluded from the processing time. Values below one indicate that processing can keep pace with the video clock under this replay measurement. Table 8 gives per-method means for the eight specialized systems. For proprietary APIs, we measure end-to-end client latency from request submission until the complete response; the full-response means are in Table 9.

Table 8: Offline replay real-time factor (RTF) for specialized memory methods. RTF is processing time to the first output token divided by the duration of the causal video prefix; each entry is the mean over evaluated probes. Values above 1 indicate that processing cannot keep pace with a live 1-FPS video stream under this replay setting. Lower is faster.

Method	Mean RTF
HERMES	0.204
ReKV	0.089
FluxMem	0.016
Flash-VStream	0.010
StreamForest	0.020
OASIS	1.223
Video-SALMONN S	0.010
VST	0.020

Table 9: Mean end-to-end latency (seconds) for proprietary models, measured from request submission to the complete response. Video memory replays prior sessions; text memory supplies saved summaries and the current causal prefix.

Model	Video memory	Text memory
Seed-2.0-Lite	34.11	11.25
Gemini 3.6 Flash	45.73	14.58

B.4 PERSISTENT STORAGE

Table 10 complements the normalized storage-per-video-hour comparison in Section 3.3 with the mean and maximum prior-session peak state across 104 trajectories. For general video models under *Video as Memory*, the mean and maximum trajectory-peak storage are 2,316.01 and 9,639.01 MiB, respectively; *Text Summary as Memory* occupies only kilobytes. Video-SALMONN S also maintains time-test-training fast weights whose mean trajectory-peak size is 8,404,992 bytes (8.02 MiB). This state is part of the model's parametric adaptation and does not grow with processed video length, so the main efficiency table counts its selected visual memory but does not add those fast weights to per-video storage.

Table 10: Peak persistent-memory size within each trajectory, after completed sessions. Mean and maximum are over 104 trajectories; units are shown in the cells.

Method	Mean peak	Maximum peak
HERMES	330.05 MiB	330.05 MiB
ReKV	19,202.80 MiB	58,185.20 MiB
FluxMem	167.31 MiB	189.96 MiB
Flash-VStream	565.20 MiB	566.25 MiB
StreamForest	3,374.53 MiB	3,867.47 MiB
OASIS	2,523.27 MiB	7,823.59 MiB
Video-SALMONN S	354.34 MiB	382.91 MiB
VST	4.1 KiB	6.8 KiB

B.5 INSTRUCTION FOLLOWING

The automated scorer tolerates format errors that it can repair deterministically. It marks a response incorrect only when no unambiguous task output can be recovered. Examples include an MCQA answer containing multiple options and an Adaptive response that repeats until the maximum output length without a recoverable decision. Tables 11 and 12 report the remaining failures. Some memory methods show more such failures with long input histories, indicating reduced instruction following under heavy context.

Table 11: Unrecoverable output-format failures for specialized memory methods. Each method is evaluated on 5,371 probes; rate is count divided by 5,371. Recoverable formatting errors are excluded.

Method	Count	Rate	Method	Count	Rate
HERMES	43	0.80%	StreamForest	1,941	36.14%
ReKV	516	9.61%	OASIS	2	0.04%
FluxMem	199	3.71%	Video-SALMONN S	792	14.75%
Flash-VStream	492	9.16%	VST	1,990	37.05%

Table 12: Unrecoverable output-format failures for general video models under video and text memory. Counts and percentages use 5,371 probes per setting. SimpleStream uses only the four most recent frames and is reported once.

Model	Video memory	Text memory
Seed-2.0-Lite	14 (0.26%)	4 (0.07%)
Gemini 3.6 Flash	4 (0.07%)	3 (0.06%)
Qwen3.8-27B	27 (0.50%)	81 (1.51%)
Qwen3-VL-8B	15 (0.28%)	9 (0.17%)
InternVL3.5-8B	98 (1.82%)	81 (1.51%)
VideoLLaMA3-7B	875 (16.29%)	816 (15.19%)
SimpleStream	0	

B.6 EFFECT OF TRAJECTORY ORGANIZATION

We compared APM-Bench trajectories with two controls on 12 EgoLife trajectories (366 candidates; 742 runtime instances). *Oracle* supplies the necessary evidence before the cutoff. *Raw Lifelong* adds the recorded video between selected sessions. All conditions use the same tasks and scoring. Figure 6 reports results for Seed-2.0-Lite, Qwen3-VL-8B, and FluxMem.

Across these three systems, APM-Bench gives a wider score spread than Raw Lifelong on Cross-session Understanding (standard deviation 8.75 vs. 3.33), Adaptive Response (8.03 vs. 4.84), and the mean of the three capability scores (8.77 vs. 7.05). We compute the population standard deviation across the three model scores as $\sigma = \sqrt{\frac{1}{3}\sum_{m=1}^{3}(s_m - \bar{s})^2}$. On Cross-session Understanding, all three systems also score higher with APM-Bench than with Raw Lifelong: 56.90 vs. 41.95, 40.05 vs. 34.16, and 37.02 vs. 35.97, respectively. Thus, the controlled trajectories make model differences in these two capabilities more visible than the longer raw history in this comparison.

Oracle raises Cross-session Understanding by 19.78–31.21 points over APM-Bench, consistent with relevant-evidence selection being an important difficulty when the history is longer. Adaptive Response gains less under Oracle, and its scores remain 28.98–51.86: access to past evidence alone does not ensure an appropriate response to the current scene. FluxMem's Real-time Perception score falls from 33.95 with Oracle to 25.14 with APM-Bench and 15.28 with Raw Lifelong, as the supplied video grows.

B.7 Evidence availability and answer accuracy

Table 13 compares Original Accuracy on all 260 questions with Evidence Available Accuracy on the 130 whose evidence remains in the two most recent sessions. Seed-2.0-Lite and Gemini 3.6 Flash score 64.62% and 85.38% on this subset, versus 62.69% and 76.15% originally. All eight specialized systems score below their original accuracy, including HERMES (7.69% vs. 38.46%). Accessible evidence alone therefore does not ensure a correct answer. Evidence Unavailable Detection evaluates the other 130 questions; Balanced Accuracy averages the two restricted-history rates.

Table 13: Memory Loss-Aware accuracy (%) on 260 questions. Original Accuracy is the accuracy on the original four-option questions under each system's main evaluation protocol, before restricting history. The restricted-history setting retains the two most recent completed sessions; Balanced Accuracy averages correct answer selection when evidence remains available and insufficient-evidence detection otherwise.

Model / Method	Original Accuracy	Evidence Available Accuracy	Evidence Unavailable Detection	Balanced Accuracy
w/o Memory				
SimpleStream	28.85	3.85	95.38	49.62
Raw Video as Memory				
Seed-2.0-Lite	62.69	64.62	66.15	65.38
Gemini 3.6 Flash	76.15	85.38	59.23	72.31
Qwen3.8-27B	58.46	59.23	46.92	53.08
Qwen3-VL-8B-Instruct	44.62	40.00	66.92	53.46
InternVL3.5-8B	36.15	50.77	40.77	45.77
VideoLLaMA3-7B	30.77	20.00	49.23	34.62
KV Cache				
HERMES	38.46	7.69	90.77	49.23
ReKV	34.23	10.77	63.85	37.31
Visual Tokens / Features				
FLUXMem	37.31	15.38	71.54	43.46
Flash-VStream	31.15	7.69	52.31	30.00
Event Tree				
StreamForest	43.85	15.38	45.38	30.38
OASIS	39.23	23.08	86.15	54.62
Parametric Memory				
Video-Salmon-S	17.31	16.15	25.38	20.77
Reasoning Thoughts				
VST	24.62	10.77	55.38	33.08

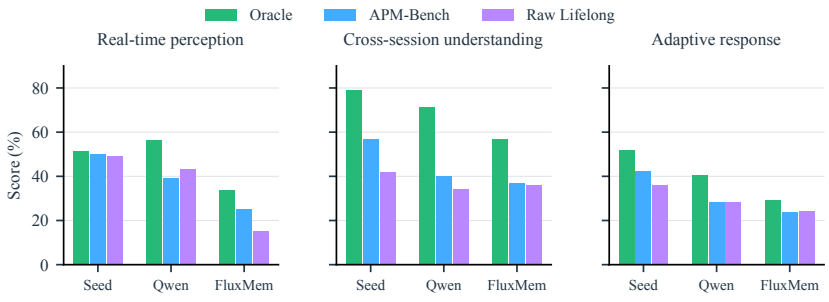

Figure 6: Trajectory organization on 366 candidates. Oracle supplies only necessary evidence; Raw Lifelong adds intervening source video. Adaptive Response uses the Section 3.3 Gated Judge score.

B.8 TASK-LEVEL PERFORMANCE

Figures 7 and 8 show the 12 task scores from Table 4. The axes group Cross-session Understanding (ER–TR), Real-time Perception (ACR–STU), and Adaptive Response (ERA–TPG). All panels use the same 0–100 scale, and both figures include SimpleStream for comparison.

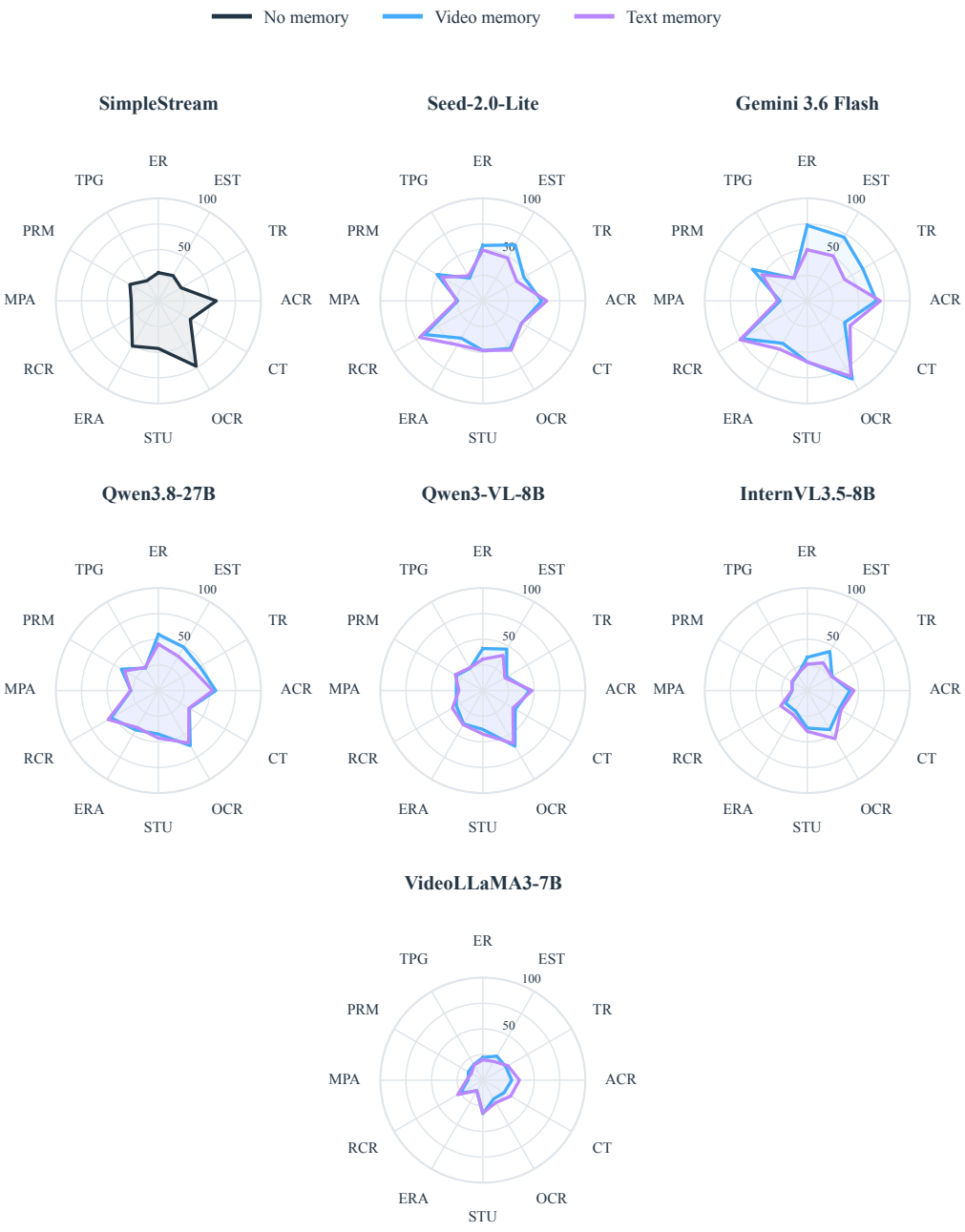

Figure 7: Scores on all 12 tasks for six general video models under video- and text-memory protocols. SimpleStream is shown in the same figure for comparison. All axes use a 0–100 scale.

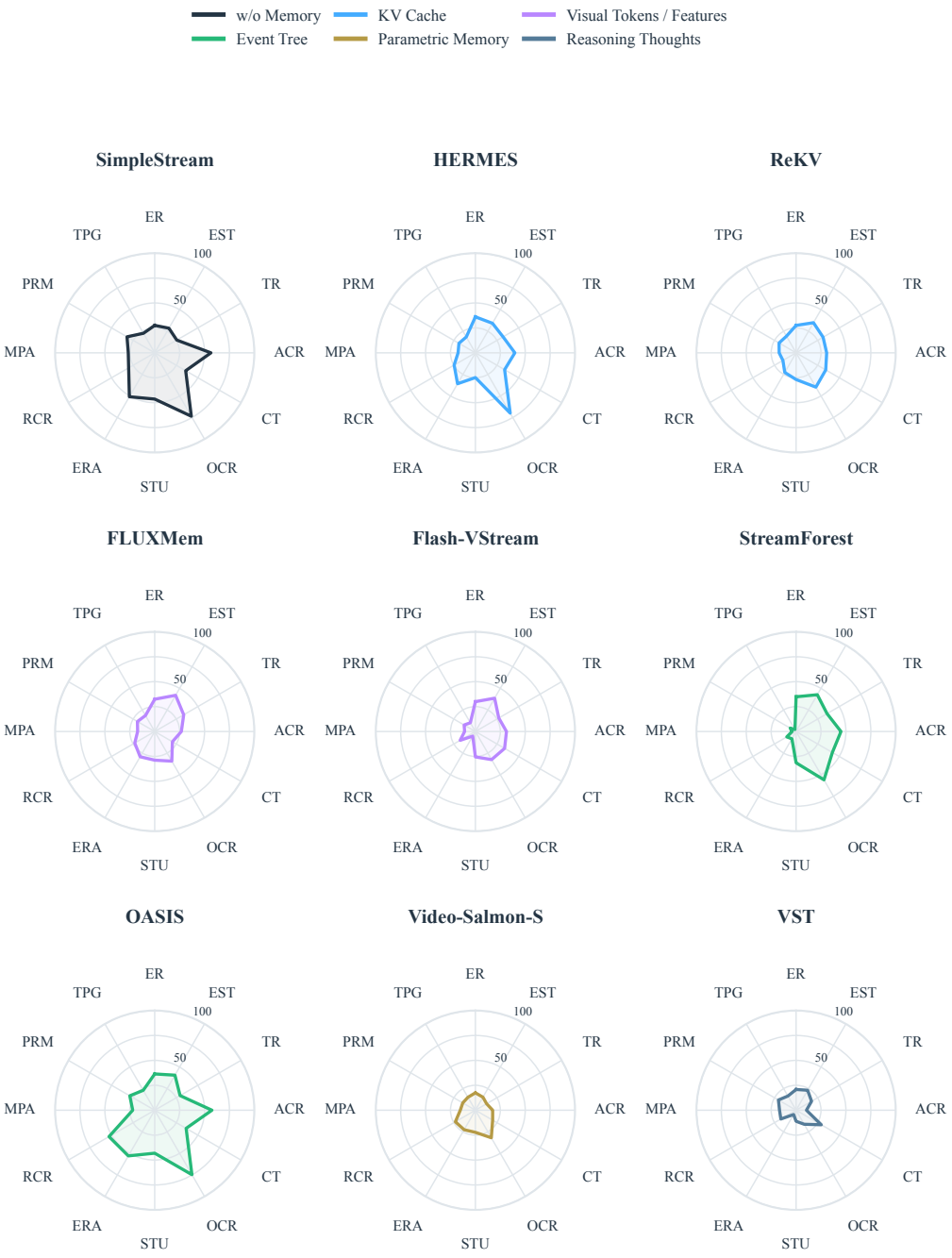

Figure 8: Scores on all 12 tasks for eight specialized streaming-memory methods and SimpleStream. All axes use a 0–100 scale.

C Dataset Statistics

C.1 Trajectory activities

The 104 trajectories cover recurring everyday activities and longer procedural tasks. Figure 9 summarizes terms from the curated EgoLife trajectory titles and the HD-EPIC recipe names. Each term is counted at most once per trajectory, so a long title does not dominate the display.



Figure 9: Activity terms in the 104 trajectory titles. Larger words occur in more trajectories.

C.2 Evidence composition

Table 14 distinguishes multi-evidence candidates from those whose evidence spans multiple sessions. Multi-evidence means that the answer or response requires at least two distinct evidence items; these items can come from the same session.

Table 14: Evidence composition for inter-session tasks with explicit evidence annotations. Multi-evidence requires at least two evidence items; multi-session evidence spans at least two sessions. Rates use the candidate count in each row. Cross-day means that at least one decisive item precedes the query day.

Task	Candidates	Multi-evidence	Multi-session	Same-day	Cross-day
ER	327	33 (10.1%)	23 (7.0%)	53.5%	46.5%
EST	298	87 (29.2%)	24 (8.1%)	55.0%	45.0%
TR	262	262 (100%)	173 (66.0%)	54.6%	45.4%
MPA	116	48 (41.4%)	19 (16.4%)	59.5%	40.5%
TPG	132	73 (55.3%)	23 (17.4%)	77.3%	22.7%

C.3 Evidence distance measured in sessions

We measure the number of session boundaries between a query and its earliest decisive evidence; for PRM, the earlier registration is the evidence anchor. Table 15 shows this distance for the 1,249 inter-session candidates. Separately, the decisive evidence occupies one, two, three, four, or five historical sessions for 987, 186, 65, 10, and 1 candidates, respectively.

Table 15: Inter-session candidates by the distance from the query session to the earliest decisive evidence session. Distance one means the immediately preceding session; the last column pools distances of four or more sessions.

Task	1 session	2 sessions	3 sessions	≥4 sessions
ER	160	80	45	42
EST	160	78	37	23
TR	69	74	50	69
MPA	63	27	12	14
PRM	91	13	3	7
TPG	85	26	8	13
All	628	298	155	168

C.4 Real-world span and organized trajectory duration

Figure 10 compares the organized trajectory duration with its elapsed real-world span, averaged by source and overall. EgoLife retains 1.3 hours of video across a mean span of 79.9 hours; the corresponding values are 0.7 and 1.9 hours for HD-EPIC and 1.1 and 58.9 hours overall. HD-EPIC already contains structured cooking procedures, so organization mainly removes segments unrelated to the recipe and separates the remaining stages into sessions. Its reduction is therefore smaller than EgoLife's.

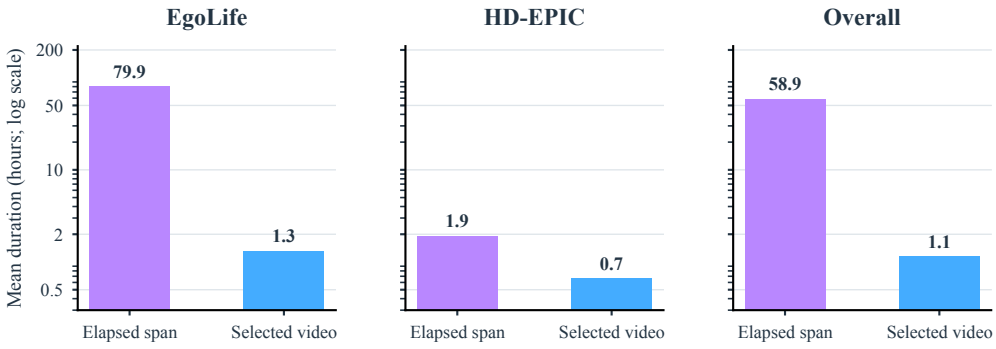

Figure 10: Mean organized trajectory duration and real-world span by source. The span runs from the first session start to the last session end, including gaps; organized duration sums the retained session videos.

D Data Construction and Annotation Details

D.1 Dataset construction

As shown in Figure 5, APM-Bench is constructed in two stages. In *Stage 1*, we organize raw videos into activity-related trajectories. For EgoLife, GPT-5-mini generates hierarchical summaries at hourly and daily scales, and GPT-5.4 mines 165 trajectory proposals from the daily summaries. Human annotators retain 76 valid trajectories comprising 472 sessions. For HD-EPIC, recipe-stage annotations yield 34 trajectory proposals, of which 28 trajectories spanning 77 sessions remain after human verification. For HD-EPIC session videos without visible timestamps, we render real-world timestamps from the source metadata onto the video frames.

In *Stage 2*, Gemini-3.5-Flash generates fine-grained timestamped visual captions for 30-second clips across all sessions. Given the full trajectory captions, DeepSeek-V4-Flash proposes Cross-session Understanding and Adaptive Response candidates; GPT-4o proposes Real-time Perception candidates from selected video frames and captions. The three capabilities yield 6,412 initial candidates. We then (1) remove shortcut candidates answered correctly without video by at least two of

Qwen3.8-27B, GPT-5, and Gemini-3.1-Pro; (2) discard candidates whose timestamps violate causal constraints; (3) use a tool-augmented VLM agent to audit and refine the remaining candidates; and (4) conduct final human verification and refinement. For 300 questions randomly sampled from the final set, two independent annotators achieved a Cohen's kappa of 0.868 (Cohen, 1960), supporting the reliability of the final annotations.

D.2 Human verification

Reviewers used the Human Verify & Refine console (Figure 11, top) to inspect and revise the 3,249 candidates retained after automated screening. The console places source video and time targets beside the review criteria and editable annotations. For MCQA, reviewers checked question clarity, answer correctness and uniqueness, answerability at query time, evidence support, timestamp accuracy, agreement between evidence descriptions and video, and absence of future leakage. Ambiguous questions were removed.

For Adaptive Response, reviewers checked each probe's intervene/silent label, whether remembered evidence helped with the current task and matched the video, whether silence was justified, and whether response and silence reasons were supported. They checked that reference responses were correct, complete, natural, and useful; verified probe and ideal-response interval timestamps, MPA/TPG historical links, RCR/PRM registrations, and the absence of future leakage; and removed unnecessary interventions or candidates with weak historical links. Reference responses and reasons were refined. The ideal intervals support finer analysis of response timing.

The independent 300-question audit used the Agreement Check console (Figure 11, bottom), which displays source evidence alongside each candidate's fields.

Each candidate is tied to a trajectory, task, and causal query point. In Adaptive Response, a positive probe marks an opportunity to respond; a negative probe marks a point at which the assistant should remain silent.

D.3 MCQA annotation format

The seven MCQA tasks (ER, EST, TR, ACR, CT, OCR, and STU) share the same four-option question and evidence structure. The five Adaptive Response annotation formats follow in Sections D.4–D.8.

MCQA: ER, EST, TR, ACR, CT, OCR, STU

```
{
  "task_type": "ER", "trajectory_id": "...",
  "candidate_id": "...",
  "question": "...",
  "choices": [{"option_id":"A","text":"..."},{"option_id":"B","text":"..."},
              {"option_id":"C","text":"..."},{"option_id":"D","text":"..."}],
  "correct_option_id": "B",
  "query": {"day_id":"DAY4","session_id":"S006",
            "query_time":"18:39:09"},
  "evidence_moments": [
    {"day_id":"DAY1","session_id":"S002",
     "start_time":"20:25:00","end_time":"20:25:30",
     "evidence_content":"..."}
  ]
}
```

D.4 ERA: evidence readiness

The four-option question is registered at session start. Subsequent probes mark when its answer becomes available. The released fields `positive_heartbeat` and `negative_heartbeats` store those probe timestamps.

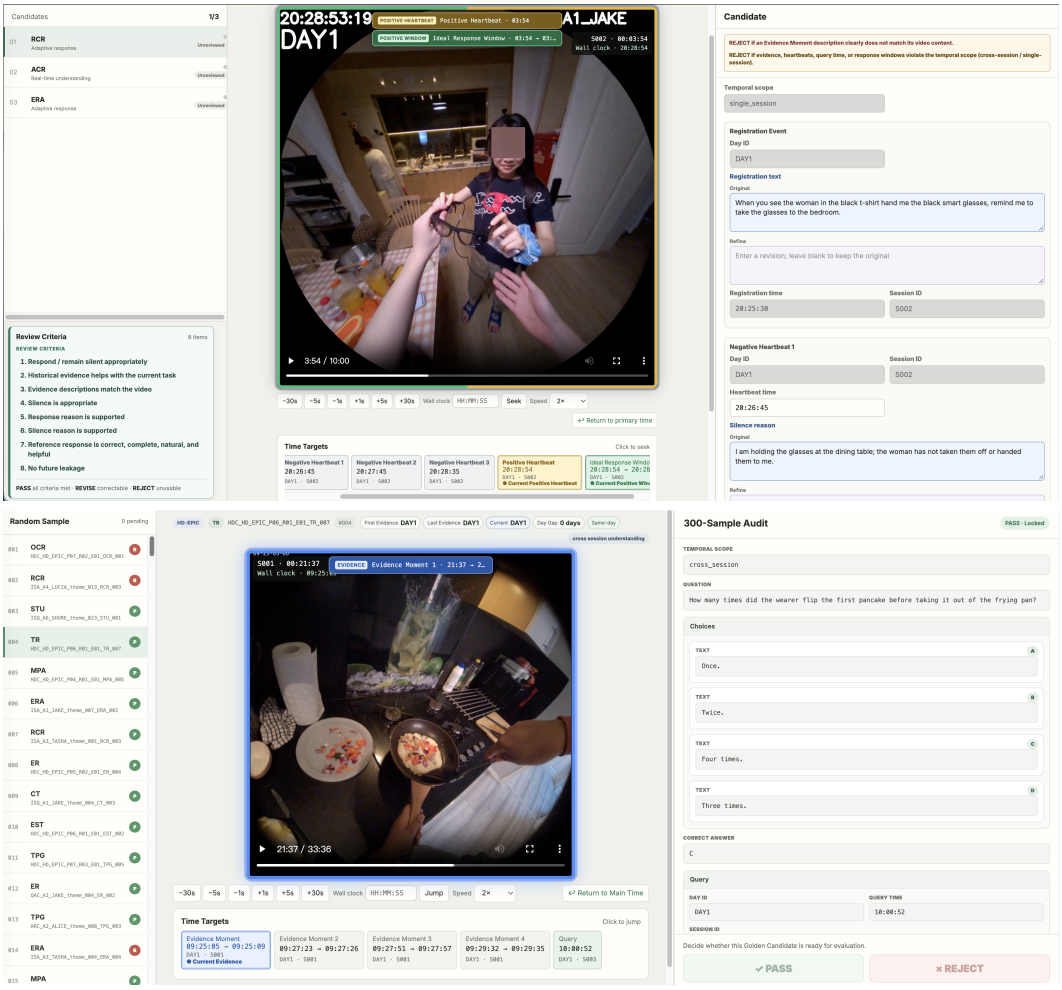

Figure 11: Human review consoles. Top: Human Verify & Refine shows video, time targets, task criteria, and editable fields for the 3,249 candidates retained after automated screening. Bottom: Agreement Check presents source evidence and candidate fields for the independent 300-question audit.

ERA: question, evidence, and probe labels


```
{
  "task_type":"ERA", "question":"...",
  "choices":[{"option_id":"A","text":"..."},{"option_id":"B","text":"..."},
             {"option_id":"C","text":"..."},{"option_id":"D","text":"..."}],
  "correct_option_id":"D",
  "answer_evidence_moments":[
    {"day_id":"DAY1","session_id":"S002",
     "start_time":"20:32:19","end_time":"20:32:27",
     "evidence_content":"..."}
  ],
  "positive_heartbeat":{"day_id":"DAY1","session_id":"S002",
    "query_time":"20:32:27","response_reason":"..."},
  "negative_heartbeats":[{"day_id":"DAY1","session_id":"S002",
    "query_time":"20:25:45","silence_reason":"..."}],
  "ideal_response_window":{"day_id":"DAY1","session_id":"S002",
    "start_time":"20:32:27","end_time":"20:32:30"}
}
```

D.5 RCR: IN-SESSION CONDITIONAL REMINDER

The user registers a condition and response in the current session. Positive and negative probes record when to fulfill the reminder or remain silent; their timestamps appear in the `heartbeat` fields.

RCR: registration and probe labels


```
{
  "task_type":"RCR",
  "registration_event":{"day_id":"DAY1","session_id":"S001",
    "registration_time":"11:12:30",
    "registration_text":"..."},
  "positive_heartbeat":{"day_id":"DAY1","session_id":"S001",
    "query_time":"11:12:56",
    "proactive_response":"...",
    "reason":"..."},
  "negative_heartbeats":[{"day_id":"DAY1","session_id":"S001",
    "query_time":"11:12:35","silence_reason":"..."}],
  "ideal_response_window":{"day_id":"DAY1","session_id":"S001",
    "start_time":"11:12:55","end_time":"11:12:57"}
}
```


D.6 MPA: ASSISTANCE FROM PRIOR EXPERIENCE

Earlier evidence supports an intervention in the current scene. The record links that evidence to a reference response and separates probes requiring a response from probes requiring silence.

MPA: historical evidence and probes; probe time = interval end_time


```
{
  "task_type":"MPA",
  "evidence_moments":[{"day_id":"DAY2","session_id":"S001",
    "start_time":"21:43:00","end_time":"21:43:03","evidence_content":"..."},
    {"day_id":"DAY2","session_id":"S001","start_time":"21:43:10",
     "end_time":"21:43:13","evidence_content":"..."}],
  "response_window_rationale":"...",
  "reference_proactive_response":"...",
  "ideal_response_window":{"day_id":"DAY7","session_id":"S003",
    "start_time":"14:24:19","end_time":"14:24:30"},
  "negative_or_silence_windows":[{"day_id":"DAY7","session_id":"S003",
    "start_time":"14:21:00","end_time":"14:21:30","silence_reason":"..."}]
}
```


D.7 PRM: CROSS-SESSION REGISTERED REMINDER

The earlier registration is paired with a later probe at which the reminder becomes due and probes at which it should remain silent.

PRM: registration and probes; probe time = interval end_time


```
{
  "task_type":"PRM",
  "registration_event":{"day_id":"DAY5","session_id":"S005",
    "registration_time":"20:55:30",
    "registration_text":"..."},
  "trigger_match_reason":"...",
  "reference_proactive_response":"...",
  "ideal_response_window":{"day_id":"DAY6","session_id":"S008",
    "start_time":"16:24:26","end_time":"16:24:27"},
  "negative_or_silence_windows":[{"day_id":"DAY6","session_id":"S008",
    "start_time":"16:23:30","end_time":"16:24:00","silence_reason":"..."}]
}
```

D.8 TPG: CONTINUING-TASK GUIDANCE

Historical evidence and the current scene identify when guidance for the continuing task is useful. The record stores the reference response and the reasons for intervening or remaining silent.

TPG: continuing-task evidence and probes; probe time = interval end_time


```
{
  "task_type":"TPG",
  "evidence_moments":[{"day_id":"DAY2","session_id":"S003",
    "start_time":"13:11:16","end_time":"13:11:23","evidence_content":"..."},
    {"day_id":"DAY2","session_id":"S005","start_time":"16:37:30",
     "end_time":"16:38:00","evidence_content":"..."}],
  "memory_relevance":"...",
  "ideal_response_windows":[{"day_id":"DAY5","session_id":"S007",
    "start_time":"23:01:28","end_time":"23:01:32",
    "response_window_rationale":"...",
    "reference_proactive_response":"..."}],
  "negative_or_silence_windows":[{"day_id":"DAY5","session_id":"S007",
    "start_time":"23:00:02","end_time":"23:00:30","silence_reason":"..."}]
}
```


E TASK PROMPTS

Angle-bracketed fields in the templates below are filled from the candidate or probe. Box titles and the two ERA dividers mark separate runtime calls; they are not part of the prompt text.

E.1 SYSTEM PROMPT

The shared instruction precedes every task prompt and restricts the assistant to information available in the causal stream.

Shared system instruction

```
You are a first-person streaming video assistant.

Use only the visual stream, persistent memory, and user interactions
    made available to you. Do not assume access to future video or to
    information that has not been provided.

Return exactly one valid JSON object in the requested format. Do not
    add markdown or text outside the JSON object.
```

E.2 MCQA PROMPTS

ER, EST, TR, ACR, CT, and OCR use the four-option template below. STU uses the same answer format with an additional instruction to interpret spatial relations from the camera wearer's viewpoint.

MCQA: ER, EST, TR, ACR, CT, OCR

```
Based only on the information available up to the current moment,
    answer the following multiple-choice question.


Question:
<QUESTION>

Choices:
A. <OPTION A>
```

```
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Return exactly:
{"answer": "A|B|C|D"}
```

STU: egocentric viewpoint guidance

```
Based only on the information available up to the current moment,
    answer the following multiple-choice question.

Task guidance:
Interpret left, right, front, behind, and other viewpoint-dependent
    directions from the camera wearer's egocentric viewpoint unless
    the question explicitly defines another reference frame. For
    object-to-object relations, use the reference object stated in
    the question.

Question:
<QUESTION>

Choices:
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Return exactly:
{"answer": "A|B|C|D"}
```

E.3 Adaptive Response probes

The five Adaptive tasks use separate probes. ERA first registers a question at session start; the two labeled parts of its box are sent at different times. RCR likewise receives the user's registration before its probe. For MPA, PRM, and TPG, an unlabeled interval starts with the marker shown below, and the probe occurs at its end time. The original template wording calls ERA and RCR probes "checkpoints" and the other probes "intervals."

ERA: question registration and response probe

```
Session-start question:
At the beginning of this session, the user asked a delayed-answer
    question. Do not guess or answer it immediately; retain it and
    wait for a response checkpoint.

Question:
<QUESTION>

Choices:
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Response probe:
This is a response checkpoint. Based only on the causal information
    available at this checkpoint, decide whether the previously asked
```

```
    question now has one reliable and uniquely determined answer. Do
    not guess or use a result that is only revealed later.

If the answer is not yet uniquely determined, return:
{"decision": "SILENT", "reason": "..."}

If the answer is now uniquely determined, return:
{"decision": "INTERVENE", "answer": "A|B|C|D", "reason": "..."}
```

RCR: response probe

```
This is a response checkpoint. Decide whether a reminder obligation
    previously registered by the user is visually triggered at this
    checkpoint and should be fulfilled now.

Use the registered condition and requested response together with the
    causal visual evidence. The existence of a registration alone is
    not a trigger. Intervene only when the registered condition is
    currently observable and satisfied.

Do not assume that a response from another evaluation call has
    already been delivered. Remain silent when the registered
    condition is not currently satisfied or the visual evidence is
    insufficient.

If no response should be provided, return:
{"decision": "SILENT", "reason": "..."}

If a registered response should be provided, return:
{"decision": "INTERVENE", "proactive_response": "...", "reason":
    "..."}
```

Interval-start marker for MPA, PRM, and TPG

```
The unlabeled evaluation interval begins now at <START_TIME>. Assess
    assistance only from this marker through the current cutoff.
```

MPA: probe at visual-interval end

```
Evaluate the following unlabeled half-open interval in the causal
    visual stream:
[<START_TIME>, <END_TIME>)

Decide whether useful and timely proactive assistance should be
    provided during that interval. Use earlier experiences,
    preferences, or repeated behavior only when they materially
    improve concrete assistance in the current situation.

Intervene only when the currently observable situation makes a
    response useful and actionable. Remain silent when the evidence
    is insufficient, the relevant condition has not been met, or the
    opportunity is not currently actionable.

If no response should be provided, return:
{"decision": "SILENT", "reason": "..."}

If a response should be provided, return:
```

```
{"decision": "INTERVENE", "proactive_response": "...", "reason":
    "..."}
```

PRM: probe at visual-interval end

```
Evaluate the following unlabeled half-open interval in the causal
    visual stream:
[<START_TIME>, <END_TIME>)

Decide whether useful and timely proactive assistance should be
    provided during that interval. Consider whether a previously
    registered reminder obligation is visually triggered during this
    interval and should be fulfilled. Use the registered condition
    and requested response together with the causal visual evidence.
    The existence of a registration alone is not a trigger. Intervene
    only when the registered condition is currently observable and
    satisfied.

Intervene only when the currently observable situation makes a
    response useful and actionable. Remain silent when the evidence
    is insufficient, the relevant condition has not been met, or the
    opportunity is not currently actionable.

If no response should be provided, return:
{"decision": "SILENT", "reason": "..."}

If a response should be provided, return:
{"decision": "INTERVENE", "proactive_response": "...", "reason":
    "..."}
```

TPG: probe at visual-interval end

```
Evaluate the following unlabeled half-open interval in the causal
    visual stream:
[<START_TIME>, <END_TIME>)

Decide whether useful and timely proactive assistance should be
    provided during that interval. Consider whether progress in the
    same ongoing longer-term task makes stage-appropriate guidance
    useful now.

Intervene only when the currently observable situation makes a
    response useful and actionable. Remain silent when the evidence
    is insufficient, the relevant condition has not been met, or the
    opportunity is not currently actionable.

If no response should be provided, return:
{"decision": "SILENT", "reason": "..."}

If a response should be provided, return:
{"decision": "INTERVENE", "proactive_response": "...", "reason":
    "..."}
```

E.4 Memory Loss-Aware prompt

The Memory Loss-Aware evaluation gives models access only to the two most recent completed sessions. Its fifth option asks the model to identify questions whose required evidence is outside that accessible history.

Memory Loss-Aware MCQA: evidence-sufficiency variant

```
Based only on the causal information available up to the current
    moment, first determine whether the evidence is sufficient to
    support one reliable answer and whether this is the appropriate
    time to answer.


If the evidence is sufficient, select the best content answer. If the
    evidence is insufficient or answering would require information
    revealed only later, select the option that explicitly states it
    is not yet appropriate to answer.

Question:
<QUESTION>

Choices:
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>
E. <OPTION E>

Return exactly:
{"answer": "A|B|C|D|E"}
```

E.5 TEXT-SUMMARY MEMORY PROMPT

At each session end, the memory writer produces a summary without access to future questions. Summaries from completed sessions are supplied with the current causal video prefix at later queries or Adaptive probes. The writer uses the following instruction and session-end request.

Text-summary protocol: memory-writer instruction

```
You maintain query-agnostic persistent memory for a first-person
    streaming video assistant.

Observe this complete session and the user interactions that occur
    during it. After the session ends, write a detailed,
    self-contained memory of information that may remain useful in
    later sessions. The original session will not be available then.

Decide what to retain and how to organize it without anticipating any
    future evaluation question. Preserve events, entities, states,
    locations, task progress, corrections, preferences, and
    registered obligations when they are visually or explicitly
    supported. Do not invent details; preserve uncertainty when
    needed.

Return exactly one valid JSON object in this format:
{"memory": "A detailed, self-contained memory of the session."}
```

Text-summary protocol: session-end request

```
Session metadata:
- day: <DAY_ID>
- session: <SESSION_ID>
- wall-clock interval: [<START_TIME>, <END_TIME>)
The session has ended. Write its persistent memory now.
```