EMBER-BENCH: BENCHMARKING CROSS-EVENT CAUSAL MEMORY IN LONG-HORIZON EMBODIED TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Lifelong physical agents must reason over extended interactions where past events continue to shape the world long after they disappear from view. Beyond recalling what happened, agents must infer how history changes the current state and constrains future actions. Yet existing embodied and video-memory benchmarks largely focus on historical retrieval and summary, leaving such history-dependent causal reasoning underexplored. We introduce **EMBER-Bench**, an egocentric benchmark for cross-event causal reasoning in long-horizon embodied tasks, for which we newly created the task design, video recording, and data annotation. It contains 189 household tasks and 699 QA pairs, spanning task progress, failure recovery, external interventions, and compound long-horizon tasks with distant dependencies and prerequisites, with fine-grained event and causal-chain annotations. EMBER-Bench evaluates reasoning in both directions: next-action prediction selects the next action from history, and causal traceback, given that action, identifies the historical event that makes it necessary. Input ablations that add action logs or privileged cause-and-consequence annotations to the video indicate which kind of historical information models fail to use. Among the 16 evaluated models, the highest overall accuracy is 61.2%, compared with a mean of 98.3% across two human evaluators. At paired decision points, correct traceback is not associated with correct next-action prediction. Adding action logs yields a gain of 1.6 points, whereas cause-and-consequence annotations yield an additional gain of 13.0 points on top of that. These results suggest that extracting causal information from past events and converting it into constraints on current actions remains a key difficulty for long-horizon embodied agents.

1 INTRODUCTION

Lifelong agents interacting with the physical world must operate in persistent environments where past events continue to shape later decisions, even across tasks. When the consequences of past events cannot be inferred from the current observation, reliable long-horizon task execution fundamentally requires memory. Recent advances in vision-language-action models (VLAs) offer one route toward this broader vision. $\pi_{0.5}$ (Physical Intelligence et al., 2025) has demonstrated its ability to perform multi-stage tasks such as kitchen and bedroom tidying in residences not covered by training. Such progress motivates a broader question: when decision-relevant information is no longer visible, can a model use past history to determine which effects persist and choose the appropriate next action?

Task duration alone does not establish memory dependence. When the current scene exposes task progress, or a familiar action sequence suggests the next step, a model may answer correctly without consulting the relevant history. Memory must track the current task state rather than merely retain past observations. This requires distinguishing causally relevant events from incidental history and tracking how their still-valid consequences constrain current actions. We focus on decision points where, given the goal and task constraints, the current observations alone are insufficient to distinguish between task states that require different actions. These requirements include tracking task progress and prerequisites, considering the ongoing consequences of self-induced anomalies and external interventions, and maintaining these dependencies across intervention activities.

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.

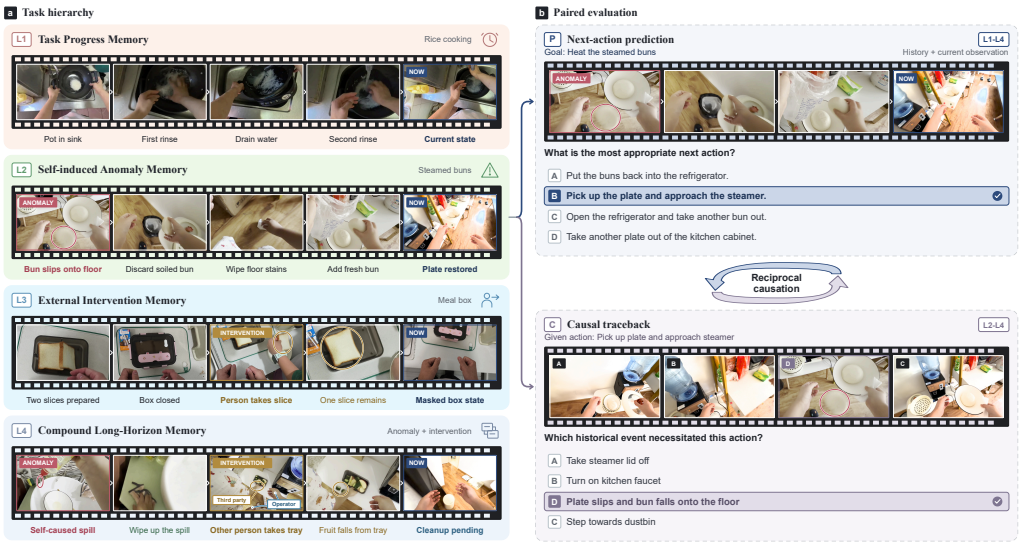


Figure 1: **Overview of EMBER-Bench. (a)** Four types of memory required for long-horizon task scenarios. **(b)** Paired next-action prediction and causal traceback assess history-grounded action selection and identification of the historical event supporting a given action.

Previous studies have examined the role of historical information from multiple perspectives. EgoSchema (Mangalam et al., 2023) and HourVideo (Chandrasegaran et al., 2024) focus on video understanding over long time spans, but long-video understanding does not by itself establish how history constrains a task-grounded action. Similarly, long embodied tasks do not necessarily rule out decisions based on current observations. EgoTaskQA (Jia et al., 2022) establishes task dependencies based on state changes and action conditions; EgoPlan-Bench2 (Qiu et al., 2026) screens the historical requirements in the next-step planning through textual priors and the current frame. FindingDory (Yadav et al., 2026) constructs tasks that rely on previous interaction experiences. Yet joint evaluation of the persistent consequences of self-induced anomalies and external interventions, together with paired action selection and causal attribution, remains limited. Existing benchmarks rarely isolate whether an embodied model remembers the causal event that explains why the current task state requires a particular action.

We introduce **EMBER-Bench**, an **E**gocentric **M**emory **B**enchmark for **E**mbodied **R**easoning over cross-event causality in long-horizon household tasks, for which we newly created the task design, video recording, and data annotation. The benchmark evaluates memory utilization through questions at fixed decision points in pre-recorded videos of human behavior, rather than directly testing the robot's closed-loop execution. Memory dependence and answer uniqueness guide construction and review: questions answerable from the goal and current frame alone are revised or excluded, and the available pre-decision history must distinguish the correct answer from the distractors. These memory requirements are reflected in four scenario types. **L1 (Task Progress Memory)** requires tracking completed steps. **L2 (Self-induced Anomaly Memory)** concerns recovery obligations arising from errors in the task performer's own actions, requiring models to select recovery actions that address lingering anomalies before normal task execution resumes. **L3 (External Intervention Memory)** concerns third-party changes to the location, identity, or availability of task-relevant objects, requiring models to update their understanding of object states and adjust current plans. **L4 (Compound Long-Horizon Memory)** concerns long-horizon tasks that combine self-induced anomalies and external interventions within the same task episode to evaluate the maintenance and updating of multiple historical constraints across intervening activities.

We construct two complementary question types: **next-action prediction** and **causal traceback**. Next-action prediction requires the model to combine the objective, current observation, and relevant history to select one of the candidate actions. For the subset of L2–L4 questions with a uniquely traceable abnormal event, causal traceback provides the reference next action and asks which histori-

cal event supports it through a still-valid consequence or task constraint. At matched decision points, we examine whether correct action-cued traceback coincides with correct next-action prediction.

Our contributions are threefold:

- We introduce **EMBER-Bench**, a newly created egocentric benchmark covering task progress, self-induced anomalies, external interventions, and compound long-horizon dependencies, with careful construction and review criteria for memory dependence and answer uniqueness.
- We pair next-action prediction with causal traceback at 151 matched decision points. Input ablations localize the key bottleneck: action history alone yields limited gains, whereas explicit causal information yields substantially larger gains.
- Across 16 models, we reveal a gap that prior benchmarks do not quantify: correct causal traceback does not necessarily translate into correct next-action prediction. Models can identify the cause given the action, yet often fail to exploit its consequences when predicting what to do next, and simple causal chain-of-thought prompting already yields initial gains.

2 Related Work

2.1 Long Video Understanding and Egocentric Memory Evaluation

Long-video understanding evaluation focuses on whether the models can locate and integrate information distributed over a relatively long time. EgoSchema (Mangalam et al., 2023), HourVideo (Chandrasegaran et al., 2024), and LongVideoBench (Wu et al., 2024) provide a basis for long-horizon context evaluation. EgoThink (Cheng et al., 2024b) offers image-level egocentric perception, reasoning and planning evaluations. In terms of historical memory, Ego4D (Grauman et al., 2022) adds episodic memory into its core tasks, and QAEgo4D (Bärmann & Waibel, 2022) builds memory QA pairs based on Ego4D and examines the utilization of historical information under restricted storage conditions. The EgoMemoria benchmark introduced by MM-Ego (Ye et al., 2025) further focuses on the perception and memory of visual details in egocentric videos of varying lengths. A longer time scope brings about the demand for information integration across events and time periods. EgoLife (Yang et al., 2025a), X-LeBench (Zhou et al., 2025), and EgoMemReason (Wang et al., 2026) have gone beyond the retrieval of a single historical fragment to study the correlation and integration of scattered information. Under continuous input conditions, TeleEgo (Yan et al., 2025), EgoSAT (Lei et al., 2026b), and EOC-Bench (Yuan et al., 2025) examine recall and prediction with history. Together, these studies examine the retrieval and integration of historical visual information as well as reasoning across events and temporal scales.

2.2 Embodied Memory and History-dependent Decision-making Evaluation

Embodied-memory evaluation spans **memory-dependent task execution** (Kwon et al., 2026; Yadav et al., 2026; Dai et al., 2026), **history-conditioned reasoning and planning** (Jia et al., 2022; Qiu et al., 2026), and **action-deviation detection and proactive assistance** (Peddi et al., 2024; Kundu et al., 2026; Gong et al., 2026) (Table 1). However, joint evaluation of how anomalies from different sources continue to constrain subsequent actions after the triggering events leave view, and whether models can trace a given action back to its supporting historical events, remains limited. To address this deficiency, **EMBER-Bench** covers both types of abnormal sources in the self-designed and collected household tasks: self-induced anomalies and external interventions. Also, **EMBER-Bench** focuses on persistent action constraints arising from self-induced anomalies and external interventions, and pairs next-action prediction with action-cued causal traceback.

3 Benchmark Design

3.1 Task Design and Data Collection

EMBER-Bench is an original diagnostic benchmark for cross-event memory in egocentric long-horizon household tasks. The tasks for this study were newly created. We also record the corresponding task videos and construct subtask annotations and QA pairs, without splicing or adapting

Table 1: Comparison with embodied-memory and long-horizon embodied-reasoning benchmarks.

Benchmark	Data	Hist. ctrl.	L1	L2	L3	L4	Focus
MEMENTO (Kwon et al., 2026)	Sim.	Task const.	–	–	–	–	Personalized object semantics and user routines.
FindingDory (Yadav et al., 2026)	Sim.	Task const.	✓	–	–	–	Conditional, ordered, multi-goal memory control.
MIKASA-Robo (Cherepanov et al., 2026)	Sim.+real	Task const.	✓	–	△	–	Object, spatial, sequence, and capacity memory.
RoboMME (Dai et al., 2026)	Sim.+real	Task const.	✓	△	△	△	Temporal, spatial, object, and procedural memory.
RMBench (Chen et al., 2026a)	Sim.+real	Task const.	✓	–	–	–	Manipulation tasks organized by memory complexity.
RoboMemArena (Lei et al., 2026a)	Sim.+real	Task const.	✓	–	△	–	Memory-driven robotic subtasks with keyframes.
EgoPlan-Bench (Chen et al., 2026b)	Reuse	Hist. prog.	✓	–	–	–	Next-action planning from goal, history, and observation.
EgoPlan-Bench2 (Qiu et al., 2026)	Reuse	Txt./frame filt.	✓	–	–	–	Screens text-only and current-frame shortcuts.
VidEgoThink (Cheng et al., 2024a)	Reuse	–	✓	△	–	–	VQA, planning, grounding, and reward modeling.
ET-Plan-Bench (Zhang et al., 2025a)	Sim.	Temp./preq.	✓	△	–	–	Preconditions, constraints, and replanning.
RoboVQA (Sermanet et al., 2024)	New	Ctxt. plan.	✓	△	–	–	Trajectories, plans, remaining steps, and success.
EmbodiedBench (Yang et al., 2025b)	Sim.	–	△	△	–	–	General embodied decision making.
HoloAssist (Wang et al., 2023)	New	–	△	△	–	–	Mistakes and human assistance in collaboration.
CaptainCook4D (Peddi et al., 2024)	New	–	△	△	–	–	Procedural errors and multistep localization.
PREGO (Flaborea et al., 2024)	Reuse	Seq. pred.	△	△	–	–	Online procedural mistake detection.
EgoProactive (Kundu et al., 2026)	New	–	✓	✓	–	–	Guidance under deviations and recovery.
ProAssist (Zhang et al., 2025b)	Reuse	Stream.	△	△	–	–	Proactive assistance timing and correction.
EgoServe / Vinci2 (Gong et al., 2026)	Reuse	Mem. horiz.	✓	✓	–	–	Assistance from immediate to cross-session context.
EMBER-Bench (Ours)	New	Hist. mask + review	✓	✓	✓	✓	Cross-event state updates and bidirectional causal QA.

Notes. Data: New / Reuse / Sim. / Sim.+real. **Hist. ctrl.**: history-control mechanism. **L1–L4**: task progress memory, self-induced anomaly memory, external intervention memory, and compound long-horizon memory combining self-induced anomalies and external interventions within a task. ✓: explicitly evaluated; △: related or partial coverage; –: not explicitly targeted.

videos, annotations or QA samples from other benchmarks. We adopt a diagnostic assessment approach (Johnson et al., 2017; Pătrăucean et al., 2023) that organizes tests around target competencies, using "memory necessity for decision-making" as a guideline throughout the design and review phases.

We target decision points where the goal and current observation are insufficient to select the next action: relevant history must resolve task progress, object states, or action prerequisites. Thus, we construct EMBER-Bench through a detailed pipeline (Figure 2).

We design four types of memory scenarios that these tasks assess, mentioned in the introduction (Figure 1). Before we capture the video, we need to make sure each task plan of L1–L3 consists of more than 10 atomic actions and more than 40 atomic actions for that of L4. These length thresholds are used only to ensure a non-trivial interaction horizon and are not treated as evidence of memory dependence; the latter is verified separately through the memory-dependence checks described in Section 3.4. The collection of videos covers the kitchen, dining room, living room and bedroom, continuously recording egocentric household tasks. After capturing the videos, we conduct a quality-control pass to ensure adequate visual clarity and verify the interacted objects as well as anomalous or intervention events captured in the recordings.

3.2 Event and Causal Dependency Annotation

Temporal action labels alone do not specify which past event supports a later recovery action. We therefore augment manual annotations of subtask boundaries, locomotion (*navigation*), and hand interactions (*left_hand* and *right_hand*) with structured causal dependencies (Figure 2(3a)). For anomaly-related samples in L2–L4, we record the event phase (*execution_stage*), the triggering event and its effects (*accident_cause* and *accident_consequence*), and the link between a recovery action and the anomaly it addresses (*obligation_related*). These annotations connect anomalous events, their persistent consequences, and subsequent recovery actions, providing traceable evidence for bidirectional question construction. The cause-and-consequence descriptions also serve as privileged inputs in our information ablation, rather than in the main video-only evaluation. Field definitions, stage-dependent rules, and worked examples of detailed annotations for anomalies are provided in Appendix A.

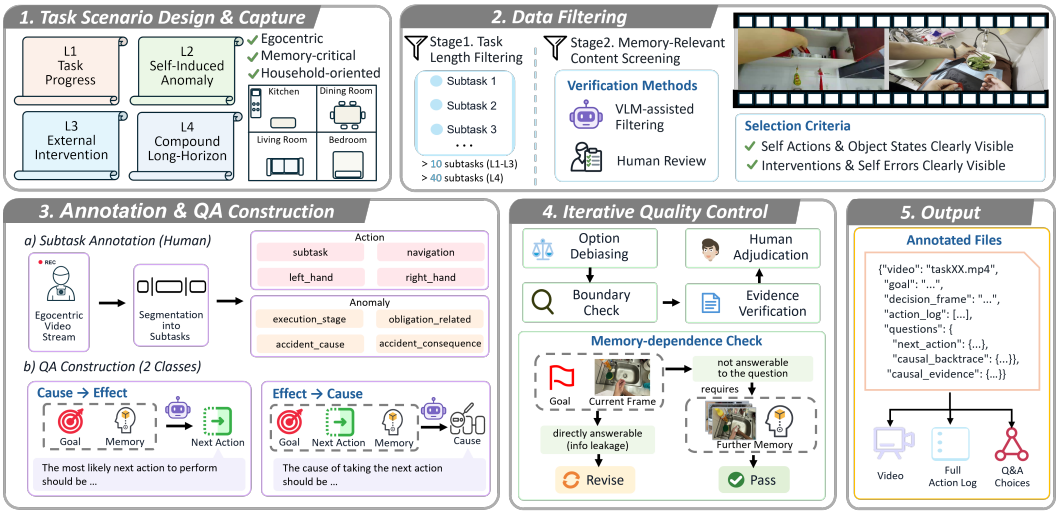


Figure 2: **The process of EMBER-Bench construction.** A human-led, model-assisted workflow covers recording, filtering, annotation, and bidirectional QA construction. Iterative quality control checks memory dependence and answer uniqueness, producing traceable benchmark samples.

3.3 Question Design and Input Protocol

Based on the event correlation annotation in Section 3.2, we transform the continuous impact of historical events on the current task state into two types of complementary causal questions: **next-action prediction** and **causal traceback** questions (Figure 2(3b)). By examining the cause-to-effect relationship, we assess whether vision-language models (VLMs) can use historical constraints to decide the next action, whereas by tracing back from the effect to the cause, we examine whether VLMs can trace the core historical basis for a given action. The problems are constructed based on the correlation of **historical events** $\rightarrow$ **event consequences** $\rightarrow$ **subsequent task conditions**. We manually construct the candidate QA pairs and answer choices, have VLMs verify them against the videos and annotations, and finally review them again manually.

First, we construct the **next-action prediction** multiple-choice questions which allow the tested models to choose one of four options. Each question has a fixed task objective and decision time (frame). The tested models are required to select the next action to be taken at that moment based on the history of previous actions and the current observation. The basis for judgment includes the confirmed status of the incident, its lingering consequences, and the prerequisites for action, which together constrain the choice of the correct answer. When constructing distractor options, we do not simply select actions that have already been completed; rather, we select actions that a model might choose if it misinterprets, ignores, or overreacts to the relevant memory. The **causal traceback** multiple-choice questions are constructed solely from the subset of next-action prediction questions that have a clear abnormal causal relationship. Some next-action prediction questions rely on the combined history formed by multiple completed steps. Although the next action can be determined, there may not be an explicit reason event suitable for a separate question, so they are retained as next-action prediction questions without being forced to be paired. For eligible L2–L4 prediction questions, we trace the annotated link from the recovery action back to the anomaly it addresses. The associated cause-and-consequence annotations identify the supporting historical event, which becomes the reference answer for the paired causal traceback question (Appendix A). In traceback questions, each option is a history clip labeled OPTION A–D in the top-left corner of the video.

The main evaluation provides the task goal, the current decision frame, a pre-decision history video sampled at 5 fps, the question, and four choices. History ends before the reference next action begins. Causal traceback additionally provides that action, with four candidate event clips marked A–D in the history video. Text-input ablations apply only to the 151 paired prediction questions (Section 4.2). Full prompts and option-rendering details are provided in Appendix B.

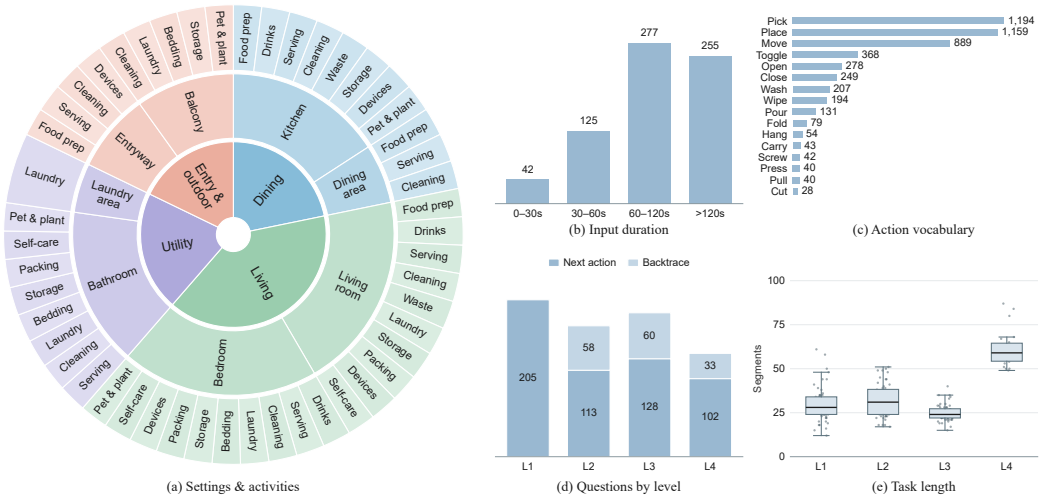


Figure 3: **Data statistics and distribution of EMBER-Bench.** **(a)** The hierarchical distribution of scenes and activities; **(b)** The distribution of video input duration based on all QA pairs; **(c)** The frequency of the 16 most commonly used types of action words; **(d)** The number of the two types of problems in L1 to L4; **(e)** The distribution of task length of L1 to L4 (segments).

3.4 Quality Control

We set the validity of the evidence, answer uniqueness, content clarity, and historical necessity as the pass criteria. First, we need to ensure that the complete video inputs provide only the observed portion (up to the start of the target next action) and do not reveal the future endpoint or subsequent results. Questions answerable from the task goal, current frame, question, and options alone are returned for revision. We conducted three rounds of testing on several mainstream VLMs, providing only the question, choices, and goal, under the condition that no model achieved full correctness across rounds. With the task goal, current frame, question, and choices but no history, accuracy is 26.49% (shown in Appendix D), compared with 25% chance accuracy. For each pass question, it should be possible to find the basis that supports the correct option and excludes the other candidates after adding the pre-decision history.

Specific reviews include checks for option debiasing, input boundary check, and memory-dependence check. We control the distribution of correct answer positions, examine the length, wording and detail granularity of the options to identify the cues, confirming that the event boundaries and input text comply with the same time limit as visual input. Through evidence verification, we reference the raw videos and action annotations to guarantee the historical facts supporting the correct answer and to check whether subsequent actions have altered the relevant circumstances. VLM assists in the initial screening, while disputed samples are checked manually based on the original video. The revised questions are re-examined, and samples that still fail to meet the requirements will be excluded.

3.5 Dataset Statistics and Distribution Analysis

EMBER-Bench contains a total of 189 self-designed and recorded household tasks, constructing 699 QA pairs, among which 548 are next-action prediction questions and 151 are causal traceback questions. The total duration of the 189 original videos is approximately 9 hours, 13 minutes, and 21 seconds. Our data production team consists of 12 members, whose work covers task design, video collection, action annotation, data filtering, QA construction, and quality verification. The dataset size reflects a selective diagnostic design. Not every long-horizon task or decision point yields a question whose uniquely supported answer requires still-relevant historical evidence beyond the goal and current frame. We therefore prioritize evidence verification, shortcut screening, and iterative revision or exclusion over maximizing task or QA counts (Section 3.4). Causal traceback is further restricted to L2–L4 cases with a uniquely traceable abnormal event (Section 3.3), so not

every next-action question produces a paired traceback question. As shown in Figure 3(a), the tasks cover living spaces such as the kitchen, dining area, and bedroom, involving household scenarios such as meal preparation, cleaning, organizing, and laundry. L1 to L4 contain 205, 171, 188 and 135 QA pairs, respectively. Next-action prediction questions cover all four scenarios, while causal traceback questions cover only L2 to L4 (numbers are shown in Figure 3(d)). Figure 3(b) shows the video input duration. Among them, there are 532 samples exceeding 60 seconds, accounting for 76.1% in total.

4 EXPERIMENTS

We evaluate 16 multimodal large language models (MLLMs; Appendix C) on four questions: dependency difficulty, the information that changes decisions, the cause–action gap, and whether explicit reasoning closes it. For video inputs, we uniformly sample the provided clips at a target rate of 5 fps. We use a common prompt per question type (Appendix B). The evaluation protocol is detailed in Appendix E.1. The main setting, denoted by V, provides only the task goal, the pre-decision history video, and the current frame.

4.1 HOW WELL DO MODELS HANDLE EACH MEMORY DEPENDENCY?

Table 2 reports accuracy by dependency level. Each question has four options (25% chance accuracy). Two human evaluators reach 98.3% overall (97.8% on P, 100% on C; Appendix E.7); the best model, Gemini 3.8 Flash, achieves 61.2% (P 58.4%, C 71.5%). Averaged over models, P accuracy is 47.1%, 35.3%, 36.6%, and 31.0% for L1–L4, respectively; 12 of the 16 models perform worst on L4. Fifteen of the 16 models score higher on C than on P; the exception, Qwen3-VL 8B, is near chance on C (25.8%). Across models, P and C accuracy correlate at $r = 0.85$, and P accuracies at different levels correlate at 0.76–0.94. The higher C accuracy may partly reflect the action cue that C provides.

For the 151 matched pairs (99 tasks), we split total history duration and the event-to-decision interval (from the end of the key event) separately into tertiles (Appendix E.3). P rises from 25.6% to 38.9% from the shortest to the longest history group, while C stays at 47.8–50.4%. As the interval grows, C falls from 59.0% to 43.5%, whereas P shows no corresponding decline (Figure 4(a–b)). In a regression controlling for dependency level and both time variables, each doubling of total history is associated with about $+11.7$ points in P, and each doubling of the interval (plus 1 s) with about -3.7 points in C; the other two associations are not significant (Appendix E.3). More distant events are therefore harder to trace, but do not necessarily correspond to lower prediction accuracy. The positive association with total duration may reflect task structure and should not be interpreted as evidence that extending the history improves decisions.

4.2 WHICH HISTORICAL INFORMATION CHANGES DECISIONS?

We progressively add two kinds of historical information to 151 prediction questions with explicit causal chains: action records and event consequences. The $V{+}L$ setting adds action descriptions, timestamps, and hand or navigation logs of subtasks completed before the decision point; the $V{+}L{+}E$ setting further adds cause and consequence annotations for completed events. Here, E is privileged information used for diagnosis; performance under this setting does not measure independent causal inference. We run this ablation on six models spanning the capability range of the leaderboard. Averaged over the six models, action logs add only 1.6 points, whereas cause-and-consequence annotations add a further 13.0 points (Figure 4(c)). The annotation gain is 18.5 points for decisions that change the original plan (71 questions) and 8.1 points for those that keep it (80 questions), a difference of about 10.4 points (Figure 4(d); Appendix E.4). For anomaly-dependent decisions, explicitly specifying how events constrain subsequent actions is thus more helpful than recovering the sequence of completed steps alone.

Annotation gains vary across models: 19.9 points for Gemini 3.8 Flash, 17.2 for Qwen3.5 397B-A17B and Qwen3.8 Omni Flash, 9.9 for Qwen3.8 27B and Qwen3.8 Flash, and only 4.0 (not significant) for Qwen3.5 122B-A10B (Appendix E.4). Given the same annotations, models differ in the degree to which they convert them into the correct action.

Table 2: Accuracy (%) under the video-history setting V. P denotes next-action prediction and C causal traceback. Models are grouped into closed-source and open-source models and sorted by overall accuracy within each group. The Human row averages two evaluators (Appendix E.7) and is excluded from ranking. The best model result in each column is shaded and bold; the second best is underlined. Unanswered items after retries are scored incorrect. Random-choice accuracy is 25%.

Model	Overall	Next-action Prediction (P)					Causal Traceback (C)			
		All	L1	L2	L3	L4	All	L2	L3	L4
Human (mean of 2)	*98.3*	*97.8*	*100.0*	*96.9*	*97.7*	*94.6*	*100.0*	*100.0*	*100.0*	*100.0*
Closed-source Models										
Gemini 3.8 Flash	**61.2**	**58.4**	**68.8**	**48.7**	**57.8**	**49.0**	**71.5**	**60.3**	**88.3**	60.6
Gemini 3.1 Pro Preview	51.5	50.7	61.0	47.8	40.6	46.1	54.3	48.3	68.3	39.4
doubao-seed-pro2.1	49.8	45.4	55.6	38.9	43.8	34.3	65.6	55.2	76.7	**63.6**
Qwen3.8 Max	47.5	44.2	52.2	39.8	44.5	32.4	59.6	56.9	68.3	48.5
GPT-6 Sol	46.2	44.2	57.1	41.6	32.8	35.3	53.6	51.7	58.3	48.5
Qwen3.8 Omni Flash	45.6	44.3	55.1	36.3	43.8	32.4	50.3	43.1	56.7	51.5
Qwen3.8 Flash	36.8	34.7	39.0	31.9	35.9	27.5	44.4	43.1	50.0	36.4
Open-source Models										
Qwen3.5 397B-A17B	45.4	43.8	52.7	42.5	38.3	34.3	51.0	48.3	61.7	36.4
Qwen3.5 35B-A3B	39.5	38.5	45.9	34.5	35.2	32.4	43.0	41.4	50.0	33.3
Qwen3.5 122B-A10B	38.2	35.8	45.4	29.2	34.4	25.5	47.0	44.8	56.7	33.3
Qwen3.8 27B	38.1	36.7	42.4	31.0	38.3	29.4	43.0	41.4	48.3	36.4
MiMo V2.6 Pro	35.8	33.4	39.5	30.1	33.6	24.5	44.4	41.4	51.7	36.4
Qwen3-VL 32B	34.5	32.5	41.0	27.4	28.9	25.5	41.7	43.1	46.7	30.3
MiMo V2.6 Flash	32.3	28.5	32.2	27.4	27.3	23.5	46.4	44.8	51.7	39.4
Qwen3-VL 8B	29.0	29.9	33.7	30.1	31.2	20.6	25.8	29.3	28.3	15.2
Qwen3-VL 235B-A22B	28.9	26.6	32.7	27.4	18.8	23.5	37.1	37.9	40.0	30.3

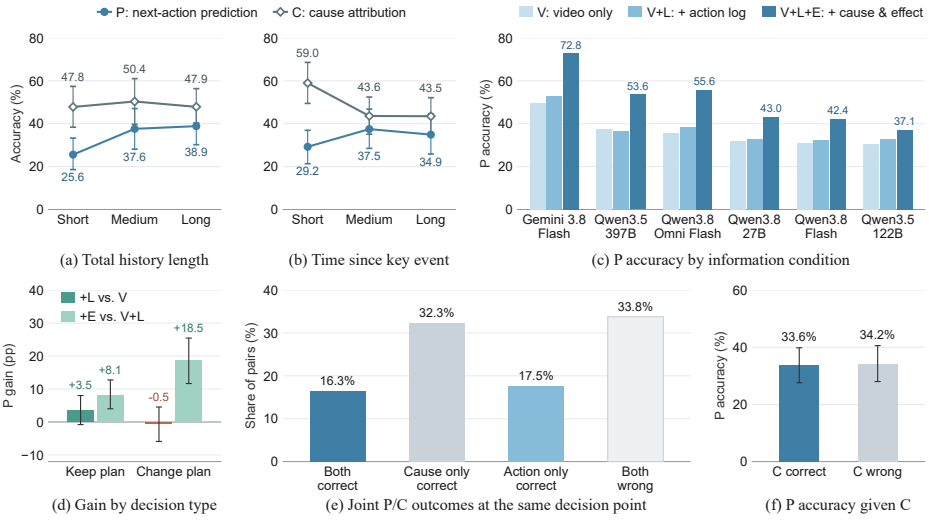


Figure 4: Temporal span, information ablation, and paired prediction–traceback outcomes. (a–b) P and C accuracy across tertiles of total history duration and event-to-decision interval. (c–d) P accuracy of six models under V, $V{+}L$, and $V{+}L{+}E$ inputs, and gains by decision type. (e–f) Joint P/C outcomes and P accuracy conditioned on C correctness. Details: Appendix E.2.

4.3 Does Identifying a Cause Coincide with the Right Action?

We pair video-only P and C responses from 16 models at the 151 matched decision points. Both answers are correct in 16.3% of the 2,416 paired responses, only C in 32.3%, only P in 17.5%, and neither in 33.8% (Figure 4(e)). This joint rate is close to the 16.5% expected under independence between P and C correctness; its low level is driven mainly by the low marginal P accuracy. P accuracy is 33.6% when C is correct and 34.2% when it is not, a difference of about -0.6 points; controlling for model gives about -2.7, and additionally controlling for question pair, which ab-

Table 3: Decision gains from causal CoT guidance. The final column is the net gain within the correct-C group (fixed by main-evaluation results), $100 \times (\text{corrected} - \text{newly wrong})/(\text{group size})$, comparing CoT against direct answering on the same 151 questions; it is not the overall direct-to-CoT change.

Model	Direct (%)	Causal CoT (%)	Net gain, correct-C group (pp)
Gemini 3.8 Flash	47.0	55.0	+13.0
Qwen3.8 Flash	31.1	34.4	+4.5
Qwen3.8 27B	31.8	27.8	0.0

sorbs item difficulty, gives about -3.5 (Appendix E.5; Figure 4(f)). None of these differences is significant, and the most adjusted estimate rules out a positive association larger than about 0.6 points. Correctly identifying the relevant event in a traceback question is thus not accompanied by a detectable gain in choosing the correct action at the same decision point. One possible explanation, consistent with this result and with the annotation gains in Section 4.2, is that models do not spontaneously exploit an event's persisting effects when selecting the next action. This interpretation requires a caveat: because C provides the action as a cue, models may solve it by matching the action to a candidate clip rather than by retrieving its cause; the results therefore establish a behavioral mismatch rather than a failure of internal causal retrieval.

4.4 Can Explicit Causal Reasoning Close the Gap?

To close this gap, we design a simple causal chain-of-thought (CoT) (Wei et al., 2022; Kojima et al., 2022) prompt that asks models to first assess whether historical events changed current conditions and whether their effects remain relevant, then select an action, without additional annotations or traceback answers (prompt in Appendix B). Because CoT substantially increases inference cost, we evaluate three models: the strongest model, Gemini 3.8 Flash, and two Qwen models of different scales (Qwen3.8 Flash and Qwen3.8 27B). The direct baseline of each model is re-measured in the same run as its CoT condition. Gemini 3.8 Flash improves from 47.0% to 55.0% (21 answers corrected, 9 newly wrong; a significant gain of about 7.9 points) and gains 13.0 points within the correct-C group (Table 3; examples in Appendix E.6). For this model, the difference in P accuracy between correct-C and incorrect-C pairs rises from about -12.3 points under direct answering to about $+5.3$ points under CoT, a significant change of about 17.6 points (Appendix E.6). Explicitly reasoning about event consequences can therefore alleviate some cases of correct traceback paired with incorrect decisions. Qwen3.8 Flash changes from 31.1% to 34.4% and Qwen3.8 27B from 31.8% to 27.8%; neither change is significant. The overall drop of Qwen3.8 27B comes from questions outside the correct-C group; within that group, corrected and newly wrong answers cancel (net 0.0 in Table 3). This pattern is consistent with a capability-dependent effect: stronger models may convert intermediate judgments about event effects into correct actions more reliably, whereas weaker models may amplify flawed judgments into confidently wrong ones. With only three models, however, this interpretation remains speculative and requires further verification; a systematic analysis of intermediate reasoning quality is left to future work.

5 Conclusion

Long-horizon embodied tasks demand more than processing longer histories: agents must identify causally relevant events and understand how their persistent consequences constrain current actions. To address the limited evaluation of this capability, we introduce EMBER-Bench, an egocentric benchmark spanning task progress, self-induced anomalies, external interventions, and compound long-horizon dependencies. Paired next-action prediction and action-cued causal traceback examine history-grounded action selection and its supporting historical events. Paired analyses show that correct traceback does not reliably coincide with correct next-action prediction; input ablations show that causal annotations, not action logs, improve decisions, and a simple causal chain of thought yields initial gains. The benchmark is limited to offline multiple-choice evaluation on pre-recorded household tasks. Future work should extend evaluation to closed-loop interaction and develop memory systems that track persistent consequences and update action constraints, supporting recovery and replanning in lifelong physical agents.

AI USE STATEMENT

Generative AI tools were used during paper preparation for language polishing, typo and grammar correction, and LaTeX formatting.

During data construction, task design, video recording, subtask annotation, and candidate QA construction were carried out by the research team. Vision-language models assisted only in verifying candidate QA pairs and in the initial quality-control screening, with reference to the raw videos and human annotations. All candidate QA pairs were subsequently reviewed by human evaluators, and disputed samples were re-examined against the original videos; samples that still failed the criteria for evidence validity, answer uniqueness, and memory dependence after revision were excluded. Model outputs were never the sole basis for determining a reference answer or retaining a sample.

The authors reviewed, revised, and verified all AI-assisted content and bear full responsibility for the study design, data quality, experimental results, and conclusions of the paper.

ETHICS STATEMENT

EMBER-Bench is intended to support research on history-dependent reasoning in embodied agents. Its household tasks, egocentric recordings, and annotations were newly created by the research team. The L3 and L4 scenarios involve filming human participants, with faces visible in some of the original recordings. Video participants agreed to participate in the study. To protect participants' privacy, we blur identifiable faces in the videos. Model evaluations are conducted offline on prerecorded videos rather than through autonomous robot operation. Benchmark performance should not be interpreted as evidence of safe or reliable deployment in real homes.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 describes task design, video collection, annotation, question construction, and quality control. Appendix A details the annotation schema and paired-question construction. Appendix B provides the evaluation prompts, input specifications for each condition, and answer-parsing rules, while Appendix D reports the current-frame-only control. Appendix C lists the evaluated models, and Appendix E describes the evaluation protocol, analysis subsets, statistical procedures, and human evaluation. These sections document how benchmark questions are constructed, presented to models, and scored, together with the procedures used for the reported analyses. The benchmark data and evaluation code are available at `https://anonymous.4open.science/r/EMBER-Bench-1A0C/`.

REFERENCES

Alibaba Cloud. Qwen3.8-Flash. Model Studio documentation, 2026a. URL `https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash`. Accessed September 25, 2026.

Alibaba Cloud. Qwen3.8-Omni-Flash. Model Studio documentation, 2026b. URL `https://www.alibabacloud.com/help/en/model-studio/qwen3-8-omni-flash`. Accessed September 25, 2026.

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. URL `https://arxiv.org/abs/2511.21631`.

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 Workshops*, pp. 1560–1568, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/Ego4D-EPIC/html/Barmann_Where_Did_I_Leave_My_Keys_-_Episodic-Memory-Based_Question_Answering_CVPRW_2022_paper.html.

ByteDance Seed Team. Seed2.1 Officially Released: Advancing AI Productivity. Official model release, June 2026. URL https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity. Accessed September 25, 2026.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-Hour Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53168–53197, 2024. doi: 10.52202/079017-1684. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5f2809607f692d79a01c05c43d702883-Abstract-Datasets_and_Benchmarks_Track.html.

Tianxing Chen, Yuran Wang, Mingleyang Li, Yan Qin, Hao Shi, Zixuan Li, Yifan Hu, Yingsheng Zhang, Kaixuan Wang, Yue Chen, Hongcheng Wang, Junjie Wang, Tianhang Yang, Renjing Xu, Ruihai Wu, Yao Mu, Yaodong Yang, Hao Dong, and Ping Luo. RMBench: Memory-Dependent Robotic Manipulation Benchmark with Insights into Policy Design. *arXiv preprint arXiv:2603.01229*, 2026a. URL https://arxiv.org/abs/2603.01229.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning. *International Journal of Computer Vision*, 134, 2026b. doi: 10.1007/s11263-025-02676-0. URL https://link.springer.com/article/10.1007/s11263-025-02676-0. Article 118.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. VidEgoThink: Assessing Egocentric Video Understanding Capabilities for Embodied AI. *arXiv preprint arXiv:2410.11623*, 2024a. URL https://arxiv.org/abs/2410.11623.

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, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_EgoThink_Evaluating_First-Person_Perspective_Thinking_Capability_of_Vision-Language_Models_CVPR_2024_paper.html.

Egor Cherepanov, Nikita Kachaev, Alexey Kovalev, and Aleksandr Panov. Memory, Benchmark & Robots: A Benchmark for Solving Complex Tasks with Reinforcement Learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1088fffae5f2aa8c8324e4f45f62248d-Abstract-Conference.html.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. RoboMME: Benchmarking and Understanding Memory for Robotic Generalist Policies. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.04639.

Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. PREGO: Online Mistake Detection in PRocedural EGOcentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18483–18492, 2024. doi: 10.1109/CVPR52733.2024.01749. URL https://openaccess.thecvf.com/

content/CVPR2024/html/Flaborea_PREGO_Online_Mistake_Detection_in_PRocedural_EGOcentric_Videos_CVPR_2024_paper.html.

Sitong Gong, 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. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.11523.

Google. Gemini 3.1 Pro Preview. Gemini API documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed September 25, 2026.

Google. Gemini 3.8 Flash. Gemini API documentation, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash. Accessed September 25, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, 2022. doi: 10.52202/068431-0242. URL https://papers.nips.cc/paper_files/paper/2022/hash/161c94a58ca25bafcaf47893e8233deb-Abstract-Datasets_and_Benchmarks.html.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are Zero-Shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Kaustav Kundu, Ritvik Shrivastava, Maxim Arap, Nanshu Wang, Xianhui Zhu, Quintin Fettes, Gautam Tiwari, Parth Suresh, Théo Moutakanni, Alejandro Castillejo Munoz, Allen Bolourchi, Pascale Fung, Pinar Donmez, Babak Damavandi, Anuj Kumar, and Seungwhan Moon. Plan, Watch, Recover: A Benchmark and Architectures for Proactive Procedural Assistance. *arXiv preprint arXiv:2606.04970*, 2026. URL https://arxiv.org/abs/2606.04970.

Taeyoon Kwon, Dongwook Choi, Hyojun Kim, Sunghwan Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. Embodied Agents Meet Personalization: Investigating Challenges and Solutions Through the Lens of Memory Utilization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.16348.

Huashuo Lei, Wenxuan Song, Huarui Zhang, Jieyuan Pei, Jiayi Chen, Haodong Yan, Han Zhao, Pengxiang Ding, Zhipeng Zhang, Lida Huang, Donglin Wang, Yan Wang, and Haoang Li. RoboMemArena: A Comprehensive and Challenging Robotic Memory Benchmark. *arXiv preprint arXiv:2605.10921*, 2026a. URL https://arxiv.org/abs/2605.10921.

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

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

OpenAI. Introducing GPT-6 Sol and Luna. Official model release, 2026. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/. Accessed September 25, 2026.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761, 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract.html.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A Dataset for Understanding Errors in Procedural Activities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135626–135679, 2024. doi: 10.52202/079017-4307. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f4a04396c2ed1342a5d8d05e94cb6101-Abstract-Datasets_and_Benchmarks_Track.html.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. EgoPlan-Bench2: A Benchmark for Multimodal Large Language Model Planning in Real-World Scenarios. *International Journal of Computer Vision*, 134, 2026. doi: 10.1007/s11263-026-02826-y. URL https://link.springer.com/article/10.1007/s11263-026-02826-y. Article 222.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. RoboVQA: Multimodal Long-Horizon Reasoning for Robotics. In *IEEE International Conference on Robotics and Automation*, pp. 645–652, 2024. doi: 10.1109/ICRA57147.2024.10610216. URL https://arxiv.org/abs/2311.00899.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_HoloAssist_an_Egocentric_Human_Interaction_Dataset_for_Interactive_AI_Assistants_ICCV_2023_paper.html.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. EgoMemReason: A Memory-Driven Reasoning Benchmark for Long-Horizon Egocentric Video Understanding. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.09874v2.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. doi: 10.52202/079017-0907. URL https://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Xiaomi MiMo Team. MiMo-V2.6: Scaling Up Reinforcement Learning for Self-Improvement. Official model release, September 2026. URL https://mimo.mi.com/docs/en-US/news/latest/v2-6. Accessed September 25, 2026.

Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. FindingDory: A Benchmark to Evaluate Memory in Embodied Agents. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2506.15635.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, Changzhi Sun, Jixiang Luo, Dell Zhang, Hao Sun, Chi Zhang, and Xuelong Li. TeleEgo: Benchmarking Egocentric AI Assistants in the Wild. *arXiv preprint arXiv:2510.23981*, 2025. URL https://arxiv.org/abs/2510.23981.

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*, pp. 28885–28900, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents. In *Proceedings of the 42nd International Conference on*

Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631, 2025b. URL https://proceedings.mlr.press/v267/yang25f.html.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MMEgo: Towards Building Egocentric Multimodal LLMs for Video QA. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b29a95e7d9f1e1a6dfc567b556733744-Abstract-Conference.html.

Yuqian Yuan, Ronghao Dang, Long Li, Wentong Li, Dian Jiao, Xin Li, Deli Zhao, Fan Wang, Wenqiao Zhang, Jun Xiao, and Yueting Zhuang. EOC-Bench: Can MLLMs Identify, Recall, and Forecast Objects in an Egocentric World? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4489. URL https://papers.neurips.cc/paper_files/paper/2025/hash/c398086709acb9938fb8c4e0868e1643-Abstract-Datasets_and_Benchmarks_Track.html.

Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, Tongtong Cao, Xingyue Quan, Yuzheng Zhuang, Yingxue Zhang, and Jianye Hao. ET-Plan-Bench: Embodied Task-Level Planning Benchmark Towards Spatial-Temporal Cognition with Foundation Models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 21566–21573, 2025a. doi: 10.1109/IROS60139.2025.11246658. URL https://arxiv.org/abs/2410.14682.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive Assistant Dialogue Generation from Streaming Egocentric Videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12044–12068. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.605. URL https://aclanthology.org/2025.emnlp-main.605/.

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 *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15206–15222. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.822. URL https://aclanthology.org/2025.findings-emnlp.822/.

A Action with Anomalies Annotation Schema

For samples involving anomalies (L2–L4), we use four labels to organize abnormal causes and consequences into causal chains, which is a key design of this benchmark. We need abnormal situations and their repaying actions to construct bidirectional memory-causal questions. Therefore, we adopt a unified structured annotation to record the correlation between event stages, obligations, abnormal causes, and consequences using four labels (shown in Table 4): *execution_stage*, *obligation_related*, *accident_cause*, and *accident_consequence*. The *execution_stage* assigns the labels of "before", "during", "repaying" or "after" to each action on the corresponding event timeline, indicating the abnormal stage it is in. The labels *accident_cause* and *accident_consequence* describe, respectively, the specific events or actions that trigger a change in an object's state during a given period and the effects of that change on ongoing task execution or the objects involved; both fields are filled in only when *execution_stage* is "during" (i.e., when the anomaly occurs during the execution of the corresponding subtask). Additionally, *obligation_related* should be filled in during the "repaying" stage, indicating the restoration or correction matters corresponding to the current processing action, containing the cause and consequences of the repaying abnormal. Except for *execution_stage*, we set all other fields to "false" in stages where they do not apply. Here "false" indicates that this field is not filled in for the current action instead of meaning that the effects of previous events have disappeared.

Table 4: Action-level annotation fields. Each action receives an execution-stage label. The other fields are populated only at the specified stage and are set to the Boolean value `false` otherwise.

Field	Guiding question	When to populate	Example values
`execution_stage`	Where does the action occur in the accident timeline?	Every action.	`before`, `during`, `repaying`, `after`.
`obligation_related`	Which accident does the action help remedy?	Only in the `repaying` stage; `false` otherwise.	"Replenish the milk" or `false`.
`accident_cause`	Which action caused the accident?	Only in the `during` stage; `false` otherwise.	"Turning around caused the insulated cup to hit the counter edge."
`accident_consequence`	What consequences did the accident cause?	Only in the `during` stage; `false` otherwise.	"Cup damage"; "Raw materials contaminated", etc.

Table 5: Example execution-stage annotations for a sequence of food-preparation actions. The `after` stage begins only when all remedial actions have been completed.

Action	`execution_stage`	Explanation
Cut cheese normally.	`before`	No accident has occurred yet.
Slice a cucumber (accidentally knocking some slices onto the floor).	`during`	The first accident occurs.
Pick up the fallen cucumber slices.	`repaying`	Attempt to address the first accident.
Throw the cucumber slices into the trash bin.	`repaying`	—
Get a cloth, wipe up the water, and put the cloth back.	`repaying`	Fulfill the floor-cleaning obligation.
Go to the refrigerator, get a fresh cucumber, and slice it again.	`repaying`	Fulfill the obligation to redo the preparation.
Resume normal food preparation after replacing the lost slices.	`after`	All remedial actions have been completed.

Here, the action sequence is just for demonstration. In our real annotations, each subtask only contains a single atomic action.

Constructing paired causal traceback questions. For eligible L2–L4 prediction questions, we start from the recovery subtask labeled *execution_stage* = "repaying" and use its *obligation_related* annotation to locate the corresponding anomalous subtask labeled *execution_stage* = "during". We then retrieve the *accident_cause* and *accident_consequence* annotations of that subtask to establish the historical basis for the reference next action. The corresponding historical event serves as the reference answer to the paired causal traceback question. Prediction questions without a uniquely traceable event remain unpaired.

B Evaluation Prompts

This appendix lists the prompt templates used in the evaluations. Fields in braces are instantiated for each question; the remaining prompt text is reproduced verbatim. Box titles and the surrounding descriptions are explanatory labels and are not part of the model input.

B.1 INPUT DESCRIPTIONS

For main next-action prediction, the following texts precede the history video and the current frame, respectively. The same video description is used in the explicit-history ablations.

Prediction video preamble

History video at 5 fps, uniformly sampled from the start up to the decision moment; then the current frame at the decision moment.

Prediction current-frame preamble

Current frame at the decision moment:

For causal traceback, the following text precedes the history video.

Causal-traceback video preamble

History video at 5 fps with OPTION A–D labels, ending before the known action; then the current frame at the decision moment.

B.2 MAIN EVALUATION: VIDEO INPUT (V)

The main next-action prediction (P) and causal-traceback (C) templates are shown below. For causal traceback, OPTION A–D labels are overlaid on the history video to identify candidate intervals; the prompt does not describe the contents of these intervals.

Prompt B.1: Next-action prediction (V, P)

Task instruction: {`objective`}
Question: What is the most likely next action to perform?
A. {`option_A`}
B. {`option_B`}
C. {`option_C`}
D. {`option_D`}

The history video is a uniformly sampled visual prefix up to the decision moment, followed by the current decision frame. Use both the video and the current frame.
If an action log is provided, it contains only completed subtasks before the decision and does not include privileged anomaly fields.
Do not write a chain of thought, explanation, or markdown.
Output exactly one line and nothing else:
Final Answer: [A/B/C/D]

Prompt B.2: Causal traceback (V, C)

Task instruction: {`objective`}
The next action is: "{`known_action`}"
Question: Which marked historical segment contains the core incident that made this next action necessary?
A. The event shown while OPTION A appears in the history video.
B. The event shown while OPTION B appears in the history video.
C. The event shown while OPTION C appears in the history video.
D. The event shown while OPTION D appears in the history video.

Each on-screen label OPTION A–D identifies one candidate historical interval.

Prompt B.2: Causal traceback (V, C) (continued)

Select the marked segment that contains the core incident that made this next action necessary.
Do not write a chain of thought, explanation, or markdown.
Output exactly one line and nothing else:
Final Answer: [A/B/C/D]

B.3 EXPLICIT-HISTORY ABLATIONS

These ablations use 151 next-action prediction questions. The V condition reuses the main-evaluation answers. Both ablation conditions use the prediction video preamble above.

The field {`action_log_json`} is a JSON list of subtasks completed before the decision. Its fields are `subtask`, `action`, `start_frame`, `stop_frame`, `start_sec`, `stop_sec`, `same_object`, `left_hand`, `right_hand`, and `navigation`. It excludes anomaly-stage, obligation, cause, and consequence fields.

Prompt B.3: Action log ($V+L$)

Task instruction: {`objective`}
Question: What is the most likely next action to perform?
A. {`option_A`}
B. {`option_B`}
C. {`option_C`}
D. {`option_D`}

You are given the complete non-privileged action log of completed subtasks up to the current decision frame. Anomaly stage, obligation, cause, and consequence fields are omitted.
action_log_nonprivileged:
{`action_log_json`}

The history video is a uniformly sampled visual prefix up to the decision moment, followed by the current decision frame. Use both the video and the current frame.
If an action log is provided, it contains only completed subtasks before the decision and does not include privileged anomaly fields.
Do not write a chain of thought, explanation, or markdown.
Output exactly one line and nothing else:
Final Answer: [A/B/C/D]

Prompt B.4: Action log and incident annotations ($V+L+E$)

Task instruction: {`objective`}
Question: What is the most likely next action to perform?
A. {`option_A`}
B. {`option_B`}
C. {`option_C`}
D. {`option_D`}

You are given the complete non-privileged action log of completed subtasks up to the current decision frame. Anomaly stage and obligation fields are omitted. Annotated accident cause and consequence for incidents in these completed subtasks are listed after the log.
action_log_nonprivileged:
{`action_log_json`}

incident_annotations (accident cause and consequence):
[

Prompt B.4: Action log and incident annotations ($V+L+E$) (continued)

```
  {
    "subtask": "{k}",
    "action": "{subtask_action}",
    "accident_cause": "{accident_cause}",
    "accident_consequence": "{accident_consequence}"
  }
]
```

The history video is a uniformly sampled visual prefix up to the decision moment, followed by the current decision frame. Use both the video and the current frame.
The action log contains only completed subtasks before the decision. Incident annotations (accident cause and consequence) refer only to those completed subtasks; they do not state the next action.
Do not write a chain of thought, explanation, or markdown.
Output exactly one line and nothing else:
Final Answer: [A/B/C/D]

Each annotated incident occupies one list item. Missing fields are filled with `(not annotated)`. When there are no annotated incidents before the decision, the incident-annotation block is replaced by the following text.

Replacement for an empty incident-annotation list

incident_annotations (accident cause and consequence):
[] (no incident was annotated in the completed subtasks)

B.4 Simple causal chain-of-thought guidance

This experiment uses 151 next-action prediction questions. The direct-answering control uses Prompt B.1. The CoT condition replaces its final three output-instruction lines with the assessment and two-line output requirements below; the complete resulting template is provided for clarity.

Prompt B.5: Causal CoT for next-action prediction

Task instruction: {objective}
Question: What is the most likely next action to perform?
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

The history video is a uniformly sampled visual prefix up to the decision moment, followed by the current decision frame. Use both the video and the current frame.
If an action log is provided, it contains only completed subtasks before the decision and does not include privileged anomaly fields.
Before choosing, briefly state whether any observed event disrupted task execution and what effect, if any, still matters for the current decision. Do not assume a disruption occurred or remains unresolved. Then select the next action.
Output exactly two lines, without markdown:
Assessment: <at most two sentences and 60 words>
Final Answer: [A/B/C/D]

B.5 CURRENT-FRAME-ONLY CONTROL

This control uses 548 next-action prediction questions. No history video or video preamble is provided; only the decision frame and the following text are sent to the model.

Prompt B.6: Current-frame-only next-action prediction

Task instruction: {`objective`}
Question: What is the most likely next action to perform?
A. {`option_A`}
B. {`option_B`}
C. {`option_C`}
D. {`option_D`}

Choose the best next action based on the task and the current image.
Output exactly one line: Final Answer: [A/B/C/D]

B.6 ANSWER PARSING

All evaluations use the same parsing rule. The answer letter following `Final Answer:` is extracted, allowing square brackets, bold formatting, or heading markers. If this marker occurs more than once with inconsistent answer letters, the response is invalid. If the marker is absent, a response is accepted only when its final line contains a single letter from A to D. A question is counted as incorrect if no valid answer is obtained after technical retries.

C BASELINE MODELS

We evaluate 16 multimodal large language models (MLLMs) on EMBER-Bench, covering seven closed-source models and nine open-source models. Closed-source models are accessible only via APIs, whereas open-source models publicly release their weights.

Closed-source models. We evaluate Gemini 3.8 Flash (Google, 2026b); Gemini 3.1 Pro Preview (Google, 2026a); Doubao-Seed-2.1-Pro (ByteDance Seed Team, 2026); Qwen3.8 Max (Qwen Team, 2026b); GPT-6 Sol (OpenAI, 2026); Qwen3.8 Omni Flash (Alibaba Cloud, 2026b); and Qwen3.8 Flash (Alibaba Cloud, 2026a).

Open-source models. We evaluate three Qwen3.5 variants: Qwen3.5 397B-A17B, Qwen3.5 35B-A3B, and Qwen3.5 122B-A10B (Qwen Team, 2026a). We additionally evaluate Qwen3.8 27B (Qwen Team, 2026b); MiMo V2.6 Pro and MiMo V2.6 Flash (Xiaomi MiMo Team, 2026); and Qwen3-VL 235B-A22B, Qwen3-VL 32B, and Qwen3-VL 8B (Bai et al., 2025).

API model versions. Some API evaluations used model aliases without an exposed immutable snapshot. Their results therefore correspond to the service versions available during the reported evaluation periods.

D VALIDATION TEST FOR NEXT-ACTION PREDICTION QUESTIONS

Table 6: Model accuracy under the setting with goal, question, options, and the current frame.

Model	Correct / 548	Accuracy (%)
Gemini 3.8 Flash	**166**	**30.29**
Gemini 3.1 Pro	147	26.82
Qwen3.8 Omni Flash	146	26.64
Qwen3.8 Max	146	26.64
Qwen3-VL 32B	145	26.46
Qwen3.8 Flash	143	26.09
MiMo V2.6 Pro	144	26.28
Qwen3.5 122B-A10B	141	25.73
Qwen3.5 397B-A17B	143	26.09
Qwen3.8 27B	142	25.91
Qwen3-VL 8B	134	24.45
Qwen3.5 35B-A3B	152	27.74
Qwen3-VL 235B-A22B	145	26.46
GPT-6 Sol	141	25.73
MiMo V2.6 Flash	138	25.18
Doubao-Seed-2.1-Pro	150	27.37
Mean (16 models)	145.19	26.49

E ADDITIONAL EXPERIMENTAL DETAILS

E.1 EVALUATION PROTOCOL

All 16 models follow the same evaluation protocol:

- **Calls.** Each question is answered in a single stateless, single-turn API call; no context is shared across questions. The temperature is set to 0 when the API supports it.
- **Reasoning.** Reasoning is disabled when the API supports it; otherwise, the lowest level is used.
- **Video input.** The history video is uniformly sampled at 5 Hz from the start of the task to the decision time and is followed by the current decision frame. For models with smaller frame budgets, the sequence is proportionally downsampled so that the entire history fits without temporal truncation; every model thus sees an evenly spaced sequence covering the full decision history. No audio is used. Reported results reflect each model's best achievable performance under its own interface constraints.
- **Prompts.** All models use the same prompts for each question type; the full prompts are given in Appendix B.
- **Retries and scoring.** Failed requests are retried; questions without a valid answer after retries are scored incorrect.

E.2 ANALYSIS SAMPLES AND STATISTICAL METHOD

Table 7 lists the models and questions used in each analysis. Each C question shares its task and decision frame with one P question, which yields the 151 matched pairs.

Table 7: Samples used in each analysis.

Analysis	Models	Questions / responses
Main leaderboard (Table 2)	16	699 per model (548 P, 151 C)
Temporal analysis (Figure 4(a–b))	16	151 matched P–C pairs
Information ablation (Figure 4(c–d))	6	151 P questions $\times$ 3 inputs
Joint analysis (Figure 4(e–f))	16	151 matched pairs; 2,416 paired responses
Causal CoT (Table 3)	3	The same 151 P questions

All confidence intervals use a task-clustered bootstrap: tasks are resampled with replacement, all questions and model responses of a sampled task are drawn together, and the interval is given by the 2.5th and 97.5th percentiles of the bootstrap distribution.

E.3 Temporal Analysis

Total history duration is the time from the start of the video to the decision frame; the event-to-decision interval is the time from the end of the correct clip of the C question to the decision frame. Each variable is split into tertiles over the 151 pairs, and tied values at a boundary are assigned to the same group (Table 8).

Table 8: Observed ranges of the temporal tertiles.

Variable	Group	Range (s)
Total history duration	Short	21.17–83.50
	Medium	84.57–134.63
	Long	136.60–594.80
Event-to-decision interval	Short	0–7.97
	Medium	8.03–30.37
	Long	30.57–163.03

The supplementary regression takes question-level accuracy as the dependent variable and jointly includes $\log_2(\text{total history duration})$, $\log_2(\text{interval} + 1)$, and level indicators. Each coefficient is the change in accuracy (percentage points) per doubling of the variable; for the interval, a doubling refers to the interval plus 1 s. The results indicate associations only and are not interpreted causally (Table 9).

Table 9: Temporal associations after controlling for level.

Outcome	Predictor (per doubling)	Coefficient (pp)	95% CI
P	Total history duration	$+11.7$	$[+5.5, +17.5]$
P	Interval + 1 s	$+1.7$	$[-0.6, +4.4]$
C	Total history duration	$+3.9$	$[-4.0, +11.9]$
C	Interval + 1 s	-3.7	$[-6.7, -0.9]$

E.4 Information Ablation

The three inputs differ only in the text appended after the question; the visual input is identical. $V{+}L$ appends the log of subtasks completed before the decision point. $V{+}L{+}E$ appends the same log followed by the `accident_cause` and `accident_consequence` of completed anomalous subtasks. Table 10 shows an example for one question.

Table 10: Appended input for one question (`Long-Horizon_L2_task3_q1`; excerpt of subtask 18).

Condition	Appended input (excerpt)
$V{+}L$	`{"subtask": 18, "action": "The arm hits the clean apple on the table", "start_sec": 80.6, "stop_sec": 82.067, "left_hand": false, "right_hand": false, "navigation": false}`
$V{+}L{+}E$	The same log, plus `{"subtask": 18, "action": "The arm hits the clean apple on the table", "accident_cause": "The arm touches a clean apple on the tabletop.", "accident_consequence": "The apple rolled onto the ground."}`

The actual input contains all logs completed before the decision point. $\Delta L = (V{+}L) - V$ and $\Delta E = (V{+}L{+}E) - (V{+}L)$ are both computed on the same questions (Table 11).

Table 11: P accuracy under the three inputs and paired gains.

Model	V (%)	$V{+}L$ (%)	$V{+}L{+}E$ (%)	ΔL (pp, 95% CI)	ΔE (pp, 95% CI)
Gemini 3.8 Flash	49.7	53.0	72.8	$+3.3$ [-4.6, $+11.3$]	$+19.9$ [$+12.9$, $+27.2$]
Qwen3.5 397B-A17B	37.7	36.4	53.6	-1.3 [-7.2, $+4.3$]	$+17.2$ [$+10.2$, $+24.5$]
Qwen3.8 Omni Flash	35.8	38.4	55.6	$+2.6$ [-4.0, $+9.4$]	$+17.2$ [$+11.0$, $+23.7$]
Qwen3.8 27B	31.8	33.1	43.0	$+1.3$ [-4.2, $+7.0$]	$+9.9$ [$+3.9$, $+16.1$]
Qwen3.8 Flash	31.1	32.5	42.4	$+1.3$ [-4.7, $+7.6$]	$+9.9$ [$+2.7$, $+17.2$]
Qwen3.5 122B-A10B	30.5	33.1	37.1	$+2.6$ [-2.1, $+7.4$]	$+4.0$ [-1.2, $+9.3$]

The decision types in Figure 4(d) are determined by the stage transition from the completed subtasks to the correct next step:

- **Change plan** (71 questions): entering recovery, exiting recovery, or responding to an anomaly that has just occurred, including a new anomaly during recovery.
- **Keep plan** (80 questions): continuing along the main task line, or remaining within a recovery process.

Every question falls into one of the two types. The annotation gain for change-plan decisions exceeds that for keep-plan decisions by $+10.4$ points ([$+2.4$, $+18.4$]).

E.5 Joint Analysis

Independence expectation. If correctness on P and C were independent, the expected proportion of pairs with both answers correct would be the product of the marginal accuracies: $33.9\% \times 48.7\% = 16.5\%$ (95% CI [13.4, 19.7]). The observed proportion is 16.3% ([12.7, 20.2]), differing from the expectation by -0.2 points ([-1.9, $+1.6$]).

Model-adjusted difference. A linear probability model is fitted to the 2,416 paired responses, with P correctness as the dependent variable and C correctness and model fixed effects as regressors. The coefficient on C is -2.7 points ([-10.04, $+4.73$]); without controlling for model, the difference is -0.6 points ([-7.60, $+6.43$]). Adding question-pair fixed effects, which absorb item difficulty, gives -3.5 points ([-7.57, $+0.58$]; 2,000 task-clustered replicates).

E.6 Causal CoT

Direct answering and causal CoT are evaluated in the same run on the same 151 questions; the prompts are given in Appendix B. The direct accuracy of Gemini 3.8 Flash in this run (47.0%) differs slightly from its main-evaluation V accuracy on the same questions (49.7%, Table 11), as its API is nondeterministic even at temperature 0, whereas the two Qwen models reproduce their main-evaluation results; Tables 3 and 12 use the same-run results throughout. The C groups follow

the C results of the main evaluation. In Table 12, *corrected* means incorrect under direct answering and correct under CoT, and *newly wrong* means correct under direct answering and incorrect under CoT. The net gain is $100 \times (\text{corrected} - \text{newly wrong})/\text{number of questions}$.

Table 12: Answer changes under causal CoT.

	All 151 questions		C-correct group	
Model	Corrected / newly wrong	Net gain (pp, 95% CI)	Questions	Corrected / newly wrong
Gemini 3.8 Flash	21 / 9	$+7.9$ [$+1.4$, $+14.7$]	108	19 / 5
Qwen3.8 Flash	18 / 13	$+3.3$ [-3.6, $+10.6$]	67	7 / 4
Qwen3.8 27B	13 / 19	-4.0 [-12.1, $+3.6$]	65	8 / 8

We also repeat the joint analysis of Section 4.3 under both prompts. The difference in P accuracy between correct-C and incorrect-C pairs changes from -12.3 ([-30.8, $+7.0$]) to $+5.3$ ([-13.2, $+23.2$]) points for Gemini 3.8 Flash (change $+17.6$, [$+5.6$, $+31.1$]), from -5.0 to -2.9 for Qwen3.8 Flash (change $+2.1$, [-12.5, $+16.4$]), and from -1.8 to $+5.2$ for Qwen3.8 27B (change $+7.0$, [-6.7, $+22.3$]). Intervals use the task-clustered bootstrap with 2,000 replicates.

Table 13 shows two wrong-to-right responses per model, one for a self-induced anomaly (L2) and one for an external intervention (L3). In the selected examples, the CoT description of the event agrees with the annotation. Model outputs are reproduced verbatim; the option text of each final answer is given in parentheses.

Table 13: Wrong-to-right examples under causal CoT.

Model	Task and key event	Direct answer (incorrect)	CoT output (correct)
Gemini 3.8 Flash	Wipe a tablet; the wipe packet slips while being torn open and the wipes scatter on the floor (L2).	Move over to the front of the table.	"Assessment: When opening the packet of wipes, it tore aggressively and the wipes scattered onto the floor by the shelf. The user just discarded the wrapper into the trash can and now needs to collect the wipes. Final Answer: [B]" (Go over to where the wipes are scattered.)
Gemini 3.8 Flash	Prepare a fruit plate; another person takes some grapes, leaving fewer than 12 (L3).	Carry the fruit plate onward into the living room.	"Assessment: Another person took grapes from the plate while the user was wiping their hands, leaving fewer than the required 12 grapes. More grapes must be added to fulfill the instruction. Final Answer: D" (Add grapes onto the fruit plate to fill it.)
Qwen3.8 Flash	Spray anti-mite spray on bedding; the robot arm knocks over the balcony trash can and garbage scatters (L2).	Carry on straightening the quilt out for drying.	"Assessment: The task of spraying the bedding is complete, and the person has moved to hang the quilt on a rack. A black trash can has fallen over with garbage scattered on the floor, which is an obstacle in the current path that needs to be addressed before continuing. Final Answer: C" (Step over to where the trash can fell.)
Qwen3.8 Flash	Boil a kettle of water; the kettle is switched off and moved onto the counter before the water boils (L3).	Raise the kettle lid without lifting the kettle.	"Assessment: The kettle was just moved off the heating base to the right side of the counter, so it is no longer positioned for boiling. To resume the task, the next step is to pick up the kettle from its current location on the right. Final Answer: C" (Pick up the kettle from the counter's right.)
Qwen3.8 27B	Get a bottle of soda; the bottle had fallen earlier, and foam overflows after the cap is loosened (L2).	Turn the bottle cap to release trapped gas.	"Assessment: The bottle was opened and is now foaming over the top, with foam visible on the neck. The most logical next step to clean up this mess before proceeding is to wipe the foam off the bottle and counter. Final Answer: C" (Wipe the foam off the bottle and counter.)
Qwen3.8 27B	Clear standing water in a bathroom; another person walks in and out of the shower area, leaving wet footprints on the mopped floor (L3).	Mop the damp patch beside the bathroom cabinet.	"Assessment: A second person entered the bathroom and left new wet footprints on the floor, which disrupted the cleaning. The most logical next step is to mop up these new wet footprints one by one. Final Answer: B" (Mop up the new wet footprints one by one.)

E.7 Human Evaluation

Two evaluators holding at least a bachelor's degree independently answered all 699 questions (548 P and 151 C).

- **Presentation.** As in the model input, each question presents the history video, the current decision frame, the task goal, the question, and four options in order. For C questions, the history video carries the same `OPTION A–D` marks, and the question states the next action.
- **Answering.** After viewing, the evaluators select one option for each question. They do not discuss the questions with each other and have no access to the action annotations, causal annotations, or reference answers.

- **Scoring.** Scoring is the same as for the models. The Human row in Table 2 averages the accuracies of the two evaluators and is excluded from ranking.

Both evaluators answered all C questions correctly; their P errors fall in L2–L4, with none in L1 (Table 14). The number of questions per level is 205, 113, 128, and 102 for P (L1–L4) and 58, 60, and 33 for C (L2–L4).

Table 14: Human evaluation results.

		P accuracy (%)							
Evaluator	P errors (L1/L2/L3/L4)	L1	L2	L3	L4	All	C errors	C (%)	Overall (%)
Evaluator 1	13 (0/4/3/6)	100.0	96.5	97.7	94.1	97.6	0	100.0	98.1
Evaluator 2	11 (0/3/3/5)	100.0	97.3	97.7	95.1	98.0	0	100.0	98.4
Mean	12	100.0	96.9	97.7	94.6	97.8	0	100.0	98.3

F LIMITATIONS

While EMBER-Bench covers four household scenario types of cross-event causal memory, there are still additions from which it would benefit. For example, as we continue data collection, we will expand beyond the current 189 tasks and 699 questions to more diverse household scenes, and extend the evaluation from offline multiple-choice questions on pre-recorded videos toward closed-loop interaction. Additionally, the benchmark currently provides test data only; constructing companion training resources that are disjoint from the test set would further support research on memory mechanisms and fine-tuning. Finally, the causal CoT in Section 4.4 is a basic probe evaluated on three models, and it is unclear to what extent more refined reasoning strategies would close the cause–action gap. We plan to explore such strategies, validate them across a wider range of models, and continue updating the benchmark as new models are released.