SUPERMEMORY-VQA: AN EGOCENTRIC VISUAL QUESTION ANSWERING DATASET AND BENCHMARK FOR LONG-HORIZON MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

AI glasses present a compelling platform for AI agents to serve as personal memory assistants. To be genuinely useful, such systems must move beyond short-term video comprehension and address memory gaps that humans experience for practical, personal, or social purposes over longitudinal egocentric video streams. However, existing egocentric datasets predominantly focus on action recognition or generic QAs from short clips, measuring perceptual capabilities rather than long-horizon human memory needs. In this work, we introduce **SuperMemory-VQA**, an egocentric visual question answering (VQA) dataset for evaluating AI assistants on practical, long-horizon memory tasks. It contains **52.9 hours** of everyday activities recorded with AI glasses, including synchronized RGB video, audio transcription, eye gaze, IMU, and SLAM trajectories. Through a human-verified annotation pipeline, we construct grounded **4,853** question-answer pairs that span object and location memory, intent recall, visual scene recall, timeline reconstruction, conversational memory, and in-context retrieval. Each question is posed as multiple-choice with an explicit "unanswerable" option to test hallucination robustness. Benchmarking leading agentic frameworks and LLM backbones reveals that existing systems remain far from reliable on real-world memory tasks, highlighting the need for new architectures for grounded AI memory that can answer only when evidence is sufficient. A participant survey further supports that our questions are realistic, useful, and aligned with everyday memory needs.

1 INTRODUCTION

The integration of AI agents into AI glasses represents a transformative leap in personal computing. By continuously observing a user's daily life, AI glasses create a unique opportunity for AI agents to serve as **personal memory assistants** in the physical world, helping users recall information from everyday life, and providing support for those struggling with memory loss. While short-term visual understanding is relatively tractable because the relevant evidence is typically recent or still in view, practical memory assistance requires substantially more: retrieving information across long horizon and integrating evidence distributed over multiple observations and episodes.

Existing progress toward such personal memory assistants is limited by both model capability and evaluation datasets. Even models with million-token context windows (Team et al., 2023) degrade over long horizon and suffer from the "lost in the middle" effect (Liu et al., 2024). Retrieval-augmented approaches such as (Lewis et al., 2020) mitigate the context limit but often fail at compositional, multi-hop reasoning over long temporal gaps (Xue et al., 2025). More importantly, existing egocentric datasets predominantly target different capabilities. Many focus on action recognition, activity understanding, localization, or generic question answering over relatively short clips, measuring perceptual capabilities rather than realistic human memory needs (Damen et al., 2022; Bärmann & Waibel, 2022; Mangalam et al., 2023b; Perrett et al., 2025).

To fill this gap, we introduce **SuperMemory-VQA**, an egocentric visual question answering (VQA) dataset and benchmark for evaluating AI assistants on **practical, long-horizon memory tasks**. As illustrated in Figure 1, `SuperMemory-VQA` emphasizes four critical properties missing from existing benchmarks: comprehensive memory tasks, long-horizon context, multi-evidence reasoning,

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.

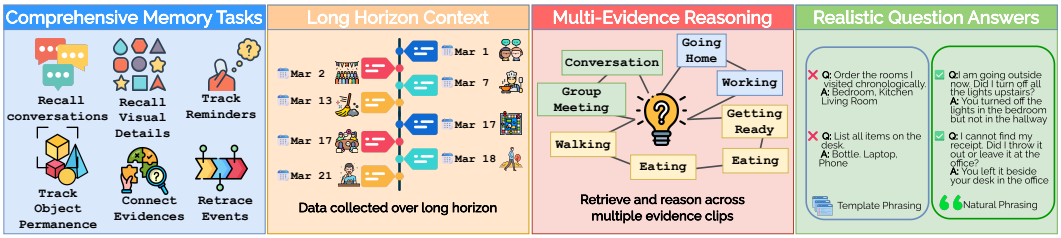


Figure 1: Overview of the SuperMemory-VQA dataset. SuperMemory-VQA advances the state-of-the-arts in four dimensions: 1) **comprehensive memory tasks**: includes six user-evaluated and commonly encountered memory tasks; 2:) **long-horizon context**: data collected over a long horizon spanning days and weeks; 3) **multi-evidence reasoning**: requires retrieving and reasoning across multiple parts of recording; and 4) **realistic question answers**: employs natural, context-grounded phrasing rather than rigid, template-based questions and answers.

and realistic question answers. Specifically, SuperMemory-VQA contains 52.9 hours of recordings of everyday activities, collected from ten participants wearing AI glasses (Engel et al., 2023) that capture synchronized egocentric RGB video, spatial audio, eye gaze, IMU, and SLAM trajectories. Each participant contributed 3–12 hours across multiple sessions, with three participants recording across multiple days up to two weeks.

Based on the recordings, we construct $4,853$ grounded QA pairs through a human-verified annotation pipeline that cover six commonly encountered memory tasks: *Object and Location Memory*, *Intent Recall*, *Visual Scene Recall*, *Timeline Reconstruction*, *Conversational Memory*, and *In-Context Retrieval*. Each question is posed as a multiple-choice item with ordered answer choices — accurate, vague, incorrect, and an explicit unanswerable option — so that safe abstention can be distinguished from confident hallucination.

We use SuperMemory-VQA to benchmark the performance of three state-of-the-art agentic frameworks – Video-RAG (Luo et al., 2024), EgoButler (Yang et al., 2025b), and WorldMM (Yeo et al., 2026) – across a diverse set of open- and closed-source Vision Language Models (VLMs) backbones, each illustrating a distinct strategy for managing long-horizon context. Our results reveal a substantial gap between current systems and reliable memory assistance on long-horizon memory tasks: they struggle with answerability detection, long temporal gaps, and evidence integration across multiple moments. These results underscore the importance of developing new techniques to fill the gap.

In summary, our main contributions are as follows:

- We introduce SuperMemory-VQA, a 52.9 hours of egocentric VQA dataset with 4,853 QAs for evaluating personal AI assistants on practical, long-horizon memory tasks.
- We develop a human-verified annotation pipeline for generating grounded QAs from egocentric data. Every QA pair is checked by a primary annotator and independently re-evaluated by a second reviewer, with identified issues adjudicated and corrected.
- We conducted a participant survey on the generated QA pairs. Our survey results support the perceived utility and practical relevance of the generated QA pairs, showing that participants view the generated QA pairs as practical, relevant, and aligned with everyday memory needs.
- Lastly, we use SuperMemory-VQA to benchmark the performance of the state-of-the-art solutions, exposing major performance gaps. We believe SuperMemory-VQA will play a significant role in developing personal memory assistants to meet practical memory needs.

2 Related Work

Egocentric Datasets. Egocentric vision captures situated human behavior. Early datasets used eye-tracking glasses for gaze and action annotations (Fathi et al., 2012; Li et al., 2015; 2018), but were small and narrow. Later efforts scaled to hundreds or thousands of hours (Damen et al., 2022; 2018; Grauman et al., 2022), though most still center on RGB video. Project Aria (Engel et al., 2023) expanded the sensor suite with synchronized RGB, eye tracking, spatial audio, IMU, and 3D scene

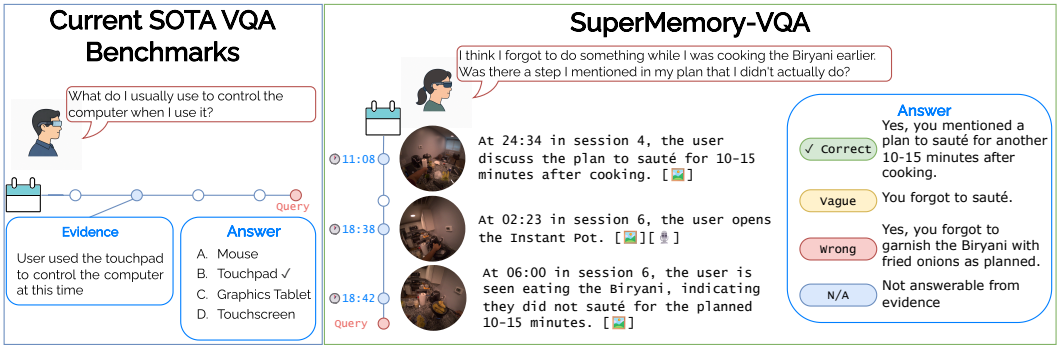


Figure 2: Comparison of SuperMemory-VQA with SOTA benchmarks. The example on the left is from EgoLife (Yang et al., 2025b): the question asks what the user usually uses to control a computer, and the answer is supported by a direct observation that the user used a touchpad at the shown time. In contrast, the SuperMemory-VQA example on the right asks whether the user skipped a planned cooking step; answering requires linking a spoken plan from one clip with later visual evidence that the user opened the Instant Pot and ate the food without the planned sauté step. This illustrates how SuperMemory-VQA emphasizes practical memory questions with long-horizon, multi-modal evidence and ordered answer choices.

context across growing egocentric recordings (Grauman et al., 2024; Ma et al., 2024; Lv et al., 2024; Perrett et al., 2025). Recent work (Yang et al., 2025a) demonstrated that gaze and other non-visual cues improve video understanding. Despite this progress, existing benchmarks do not evaluate the longitudinal, multi-session memory practical AI assistants require—recalling where objects were left across days, reconstructing daily timelines, or tracking conversational commitments. The closest work to ours is EgoLife (Yang et al., 2025b), an egocentric dataset spanning a week. Its week-long recordings and EgoLifeQA questions address practical memory needs, including recalling objects, identifying habits, understanding social interactions, and tracking tasks. SuperMemory-VQA extends this direction with two specific challenges: whether an assistant gives an answer precise enough to satisfy the question, and whether it abstains when the recording cannot support an answer.

Long Video Understanding. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023) retrieves external information beyond limited language model contexts. In video, captions often form a retrievable text corpus, while multimodal RAG indexes frames directly (Reddy et al., 2025; Wan et al., 2025; Luo et al., 2024). Video-RAG (Luo et al., 2024) jointly retrieves frames, ASR, OCR, and detections to preserve visual details, but frame-level retrieval remains hard to index and scale (Reddy et al., 2025; Wan et al., 2025). Structured variants include Video-RAG (Ren et al., 2025) for parallel text-, visual-, and graph-based clip retrieval; EgoButler (Yang et al., 2025b), which combines egocentric audiovisual clip understanding with hierarchical summaries and retrieval over long recordings; AdaVideoRAG (Xue et al., 2025) for adaptive retrieval; GraphVideoAgent (Chu et al., 2025), which extends VideoAgent (Wang et al., 2024) with caption-derived graphs; and VideoMindPalace (Huang et al., 2025), whose room-level spatial graphs limit open-ended use. WorldMM (Yeo et al., 2026) constructs episodic, semantic, and visual memories at multiple temporal scales and uses an adaptive retrieval agent that iteratively selects the memory source and granularity for each query. However, our tasks—object tracking, conversation recall, and timeline reconstruction—require compositional reasoning over entity states across sessions and heterogeneous modalities. SuperMemory-VQA provides task categories and ground truth beyond single-video QA, covering multi-day long-horizon memory.

3 SuperMemory-VQA Dataset

3.1 Overview

SuperMemory-VQA assesses whether state-of-the-art agentic frameworks can provide practical, data-driven answers to practical memory queries in everyday lives. It contains 52.9 hours of egocentric data from ten participants wearing Gen 1 Meta Aria Glasses (Engel et al., 2023); each participant



Figure 3: Example VQA pairs from each of the six task categories.

contributed 3 to 12 hours across multiple sessions of recordings, with three participants spanning multiple days (up to 2 weeks). The dataset in multimodal: it includes synchronized video, audio transcripts, eye gaze, motion, and SLAM data. Appendix 8 provides the full hardware specification, protocol, anonymization procedure, and release details.

We designed the queries to reflect what a human user would *want* to remember for practical, personal, or social purposes. The six task categories capture this scope:

- **Object & Location Memory** recalls the last known position of an object and its trajectory across time and locations, based on the concept of *episodic memory* (Tulving, 1972).
- **Conversational Memory** recalls spoken facts, commitments, deferred answers, and mid-conversation corrections. Also known as Dialogue State Tracking (DST), this is a formal component within task-oriented conversational AI (Henderson et al., 2014; Wu et al., 2019).
- **Visual Scene Recall** retrieves visual details such as text, screens, manuals, ingredients, or scene contents. While rooted in episodic memory, it concurrently evaluates the system's *semantic memory* (Tulving, 1972; Squire, 2004).
- **In-Context Retrieval** combines current context with prior facts. It evaluates the system's capacity for *relational memory* (Eichenbaum, 1997; Cohen et al., 1997)—the cognitive ability to represent and navigate associations between independent elements of an experience.
- **Timeline Reconstruction** chronologically sequences events, evaluating the temporal aspect of episodic memory, and *procedural memory* (Cohen & Squire, 1980; Squire, 2004).
- **Intent Recall** recovers stated or implied goals, reminders, and intended future actions, capturing *prospective memory* (Einstein & McDaniel, 1990; Brandimonte et al., 1996).

Figure 3 shows examples of all these tasks. We also asked participant to link their chain of thought when answering questions from their data to a set of memory strategies. Figure 5 shows this and validates the taxonomy through participant-described memory strategies. Together, these queries challenge AI systems across five dimensions:

- **Natural Phrasing.** `SuperMemory-VQA` uses conversational, context-dependent queries rather than template-style questions, requiring models to infer user intent, temporal reference, and memory context instead of relying on predictable syntax.

Table 1: Comparison of SuperMemory-VQA with existing egocentric benchmarks.

Dataset	Focus	Hrs	Context	QAs	Multi-Evid.	Natural	Evaluation
EPIC-KITCHENS-100 (Damen et al., 2022)	Action Rec.	100	≈8.5m	–	–	No	Verb-noun labels & narrations
Ego4D (Grauman et al., 2022)	Ego Activities	3,670	≈23m	–	–	No	Temporal/spatial localization labels
EgoSchema (Mangalam et al., 2023a)	Long Video QA	> 250	3m	5,063	Single	No	5-way MCQ over clips
HoloAssist (Wang et al., 2023a)	Assistance	166	≈5–10m	–	–	No	Mistakes, interventions
AEA (Lv et al., 2024)	Spatial	> 7.5	≈5m	–	–	N/A	No QA; sensor/MPS outputs
Nymeria (Ma et al., 2024)	Motion	300	≈15m	–	–	N/A	Motion/action descriptions
EgoLife (Yang et al., 2025b)	Life Assistant	300	>1h	6,000	7.2%	No	Generic MCQ w/ evidence timestamps
SuperMemory-VQA (Ours)	**Long-Horizon Memory***	52.9	**>1h**	**4,853**	**34%**	**Yes**	**Ordered MCQs with time spans**

–: no QA benchmark or QA evidence annotations. Multi-evid. means a QA requires more than one supporting timestamp or clip.
* Encompasses multiple aspects of supermemory (Appendix 9).

- **Long-Horizon Context.** Questions are grounded in recordings that last hours and sometimes span days, exceeding the practical context limits of current vision-language models (Team et al., 2023; 2024) and matching the long-tail temporal complexity of egocentric video (Grauman et al., 2022).
- **Dense Multi-Evidence Retrieval.** Many of our questions require linking sparse evidence across disjoint moments, such as a spoken plan in one session and actions performed much later, so systems need retrieval-augmented reasoning (Lewis et al., 2020) and temporal abstraction to avoid "lost-in-the-middle" failures (Liu et al., 2024).
- **Grounded Multi-modal Reasoning.** Memory assistance from egocentric data requires aligning video, audio, gaze, motion, and spatial context to track actions, object states, and user intent over time. SuperMemory-VQA also stresses efficient multi-modal evidence use: for example, Figure 2 shows how auditory cues of steam venting can localize the Instant Pot event before visual confirmation.
- **Epistemic Calibration and Hallucination Robustness.** Instead of standard independent multiple-choice questions (Hendrycks et al., 2021; Srivastava et al., 2022), SuperMemory-VQA uses ordered answer choices that distinguish correct, vague, wrong, and unanswerable responses. Inspired by reading-comprehension abstention settings (Rajpurkar et al., 2018), this design tests whether models can avoid confident hallucinations when evidence is missing (Lin et al., 2022; Li et al., 2023) and provides ranked feedback (Correct > Vague > Wrong) for alignment methods such as DPO (Ouyang et al., 2022; Rafailov et al., 2023). Together, these properties make SuperMemory-VQA a benchmark for next-generation multi-modal memory assistants in wearable glasses; Appendix 10 discusses limitations and future extensions.

3.2 Comparison to Existing Datasets

Existing egocentric datasets advance first-person activity understanding, long-form video QA, and multimodal life logging, including EPIC-KITCHENS (Damen et al., 2018), Ego4D (Grauman et al., 2022), EgoSchema (Mangalam et al., 2023a), Video-MME (Fu et al., 2025), AEA (Lv et al., 2024), Nymeria (Ma et al., 2024), and EgoLife (Yang et al., 2025b). As Table 1 shows, these datasets mainly emphasize action recognition, general video comprehension, or generic template-based QAs. In contrast, SuperMemory-VQA targets practical AR memory across episodic, semantic, procedural, prospective, spatial, temporal, and conversational memory (Tulving, 1972; 2002; Squire, 2004; Einstein & McDaniel, 1990; Brandimonte et al., 1996; Henderson et al., 2014; Wu et al., 2019). Compared with prior benchmarks, SuperMemory-VQA jointly provides six functional memory tasks, multi-session evidence spanning hours or days, and substantially more multi-evidence questions (34% versus Egolife's 7.2%). Figure 2 illustrates this shift qualitatively: unlike prior benchmarks that ask about short clips or generic events, SuperMemory-VQA asks user-centered memory questions whose answers may require linking audio, visual, and temporal evidence across longer contexts. Its ordered answer choices (correct, vague, wrong, and unanswerable/"N/A") further separate safe abstention from confident hallucination (Guo et al., 2017; Geifman & El-Yaniv, 2019; Ji et al., 2023).

3.3 Data Collection

Participants wore Gen 1 Meta Aria Glasses to capture RGB video (1408x1408, 30fps), dual SLAM streams (640x480, 30fps), eye-tracking (320x240, 60fps), and 7-channel audio (48kHz). Under an

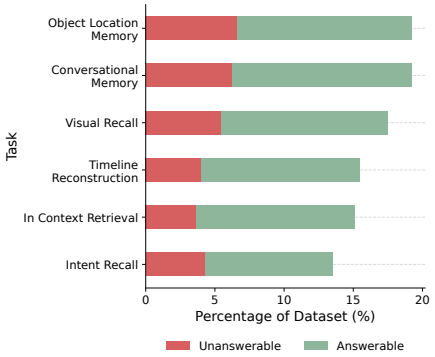


Figure 4: Distribution of answerable and unanswerable questions per task category.

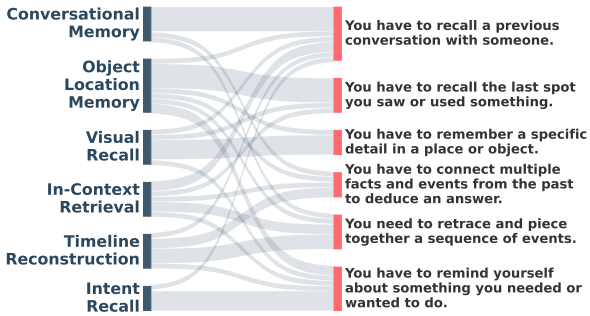


Figure 5: Sankey diagram linking task categories to participant-described memory strategies.

IRB-approved protocol, 10 participants completed guided 30–45 minute sessions in a simulated home environment, including calibration, exploration, and loosely scripted indoor and outdoor tasks using assigned pseudonyms for privacy. Participant demographics are withheld to maintain privacy. Raw audio is withheld; instead, we provide WhisperX-transcribed text (Bain et al., 2023) with sensitive information manually removed. Faces and license plates are obscured using EgoBlur (Raina et al., 2023), and direct interactions with non-participants were excised. Appendix 8.2 details the protocol, recruitment, scripts, and follow-up sessions; Appendix 8.3 details de-identification and non-participant handling.

3.4 Annotation Pipeline

We develop an automated, agentic pipeline with two phases, each followed by human review.

Phase 1: Dense Video Captioning. Video chunks and WhisperX audio transcriptions are processed by an LLM Captioning agent. Using a dynamic Person Registry, it extracts descriptions of visual actions, objects, and auditory events, aggregating them into consolidated Video Captions (Figure 9a).

Phase 2: Agentic QA Generation and Human Review Captions and metadata are merged into a unified "Super Ledger." A QA Planner proposes rationale-first (Willard & Louf, 2023), instance-level chain-of-thought (Xiao et al., 2024) QA pairs targeting the dataset dimensions and tasks (Wang et al., 2023b). A Verifier checks each pair against criteria such as factual correctness, causality, and naturalness by generating criterion-level rationales and scores (Zheng et al., 2023) while querying a Retriever over the Super Ledger; an Enhancer iteratively refines pairs. Approved pairs enter the Accepted Set for final human review (Figure 9b). The Captioning, Planner, and Verifier agents use Gemini-3.1-Pro-preview, while the Retriever and Enhancer use Gemini-3-Flash-preview. The Appendices 12, 12.5, 16.2, and 13.1 detail the phases, schema, verification, and the human review workflow.

3.5 Dataset Statistics

Our dataset includes $4,853$ QA pairs from 52.9 hours of multimodal egocentric video, balanced across task types while preserving variation in question complexity, as shown in Figure 4. Object Location Memory ($\sim$19%) and Conversational Memory ($\sim$18%) are the most common, with Visual Recall, In Context Retrieval, Timeline Reconstruction, and Intent Recall making up the rest. Appendix 13 reports temporal gaps, evidence counts, and evidence-duration distributions.

4 Experimental Setup

Baselines: We benchmark the performance of three state-of-the-art frameworks on our `SuperMemory-VQA` dataset – Video-RAG (Luo et al., 2024), EgoButler (Yang et al., 2025b),

Table 2: Performance of open- and closed-source vision-language models across agentic frameworks.

Model	Video-RAG			EgoButler			WorldMM		
	Ans-F1	Acc.	MRR	Ans-F1	Acc.	MRR	Ans-F1	Acc.	MRR
Open-source models									
Qwen-3.5 9B	65.6	46.7	66.1	53.0	40.7	61.5	66.8	46.8	66.9
Qwen-3.8 27B	69.4	51.0	68.2	46.0	41.1	61.3	71.3	52.8	70.3
InternVL-3.5 8B	81.7	41.0	63.3	61.4	39.8	61.8	59.6	36.7	59.6
Gemma-4-E4B IT	40.3	35.3	58.2	30.9	36.4	58.2	66.3	49.1	67.6
Closed-source models									
Gemini-3.8-Flash	76.0	60.9	**76.6**	74.4	**59.4**	**75.6**	**79.8**	70.5	**82.6**
Gemini-3-Flash	**83.9**	**61.0**	76.0	71.2	54.1	71.6	76.9	66.8	79.4
Gemini-3.1-Pro	67.4	53.2	70.7	43.5	42.6	64.2	72.9	67.6	80.1
GPT-6-sol	81.7	59.7	74.7	**76.8**	56.2	72.9	78.4	**70.6**	81.7

Best score in each framework–metric column is highlighted.

and WorldMM (Yeo et al., 2026) – each illustrating a distinct strategy for managing long-horizon context[1].

Video-RAG (Luo et al., 2024) is a training-free, single-turn retrieval-augmented pipeline that augments a VLM with auxiliary text extracted from the source video. Three parallel databases were precomputed per session: ASR transcripts (Whisper-X (Bain et al., 2023)), OCR text from sampled frames (EasyOCR (JaidedAI, 2020)), and object detections on CLIP-selected keyframes (APE (Shen et al., 2024)). At inference, the VLM decomposes the query into a retrieval request, each database is queried via FAISS (Johnson et al., 2019) over Contriever embeddings (Izacard et al., 2021), and the retrieved texts are concatenated with sampled frames as input to the VLM.

EgoButler (Yang et al., 2025b), proposed alongside EgoLife, is the closest in spirit to `SuperMemory-VQA` as it explicitly targets long, multi-day egocentric VQAs. It pairs EgoGPT, an omni-modal VLM that produces dense visual–audio captions with EgoRAG, which recursively summarizes these captions into hour- and day-level digests to form a hierarchical memory bank. At query time, EgoRAG performs coarse-to-fine temporal localization, retrieving high-level summaries first, then narrowing to clips and passing the top-k clip captions to the VLM for answer generation.

WorldMM (Yeo et al., 2026) builds a multimodal memory bank with three complementary components: episodic memory records events at multiple temporal scales, semantic memory updates higher-level knowledge, and visual memory retains scene and appearance details that textual summaries may omit. At inference, a retrieval agent iteratively selects the memory type and temporal granularity relevant to the question, and a response agent uses them to generate the final answer.

Base Models. We instantiate each framework with the VLMs in Table 2. Open-source models are implemented on 4×A100 GPUs; closed-source models are accessed via official APIs. All answer-generation calls use greedy decoding (temperature 0), while internal planning and captioning calls retain each framework's defaults. Each framework receives the question and four answer choices; to enforce causality, we cut the video at the question end time and provide only preceding segments. Appendix 11 gives further compute and reproducibility details.

Evaluation Metrics. We report three metrics in §5.1. Ans-F1 is the F1 score for the binary decision of whether a question is answerable from the available evidence or should receive the unanswerable option. QA-Acc is four-way multiple-choice accuracy, where only the ground-truth correct option receives credit; vague, wrong, and incorrect abstention choices are counted as incorrect. QA-MRR is the mean reciprocal rank induced by the model's ordered answer scores over the answer choices, rewarding models that rank the correct answer above vague, wrong, and unsupported alternatives.

5 Benchmarking Results

5.1 Main Results

Table 2 reports the performance of the baseline agentic systems introduced in §4. We evaluate each framework using the same set of open-source and closed-source vision-language models as the underlying reasoning model. We report three complementary metrics: Ans-F1 for distinguishing

[1]We also evaluated VideoAgent (Fan et al., 2024). However, it used substantially more tokens while having worse results compared to Video-RAG and EgoButler. We report limited results on VideoAgent in Appendix 16.1.

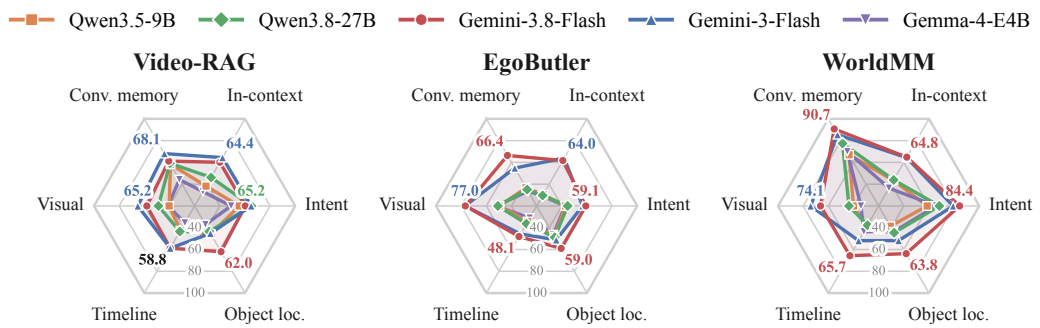


Figure 6: Task-category breakdowns across model families and agentic frameworks.

answerable from unanswerable questions, QA-Acc for selecting the correct answer, and QA-MRR for ranking the ordered answer choices. Appendix 16 reports additional results.

Overall, the results show that `SuperMemory-VQA` is challenging for both retrieval and deciding when there is enough evidence to answer. For example, Gemini-3-Flash with Video-RAG, reaches only 61.0% QA-Acc, despite achieving 83.9% Ans-F1 and 76.0% QA-MRR. This gap indicates that detecting whether a question is answerable is only the first hurdle: models must also retrieve the precise multimodal evidence, interpret it correctly, distinguish the correct answer from plausible distractors, and abstain only when the evidence is actually insufficient. These results demonstrate that long-form, real-world memory assistance remains a significant open challenge.

Across models, WorldMM outperforms both Video-RAG and EgoButler on most metrics. Among open-source models, WorldMM achieves an average accuracy of 46.4%, compared with 43.5% for Video-RAG and 39.5% for EgoButler. Among closed-source models, WorldMM achieves 68.9%, compared with 58.7% and 53.1% for Video-RAG and EgoButler, respectively. The largest gain is in Ans-F1, indicating that retrieval-augmented evidence construction is especially useful for deciding whether a question is grounded in the recorded memory.

5.2 Detailed Analysis

Closed-source models generally perform better, but the leading model varies by metric. Averaged over the four models in each group, closed-source models achieve higher Ans-F1, QA-Acc, and QA-MRR under all three frameworks. Their mean QA-Acc is 58.7% versus 43.5% for open-source models with Video-RAG, 53.1% versus 39.5% with EgoButler, and 68.9% versus 46.4% with WorldMM. No single model leads every metric. Gemini-3-Flash has the highest Video-RAG Ans-F1 (83.9%) and QA-Acc (61.0%), while Gemini-3.8-Flash leads Video-RAG QA-MRR (76.6%). Under EgoButler, GPT-6-sol leads Ans-F1 (76.8%), whereas Gemini-3.8-Flash leads QA-Acc (59.4%) and QA-MRR (75.6%). Under WorldMM, Gemini-3.8-Flash leads Ans-F1 (79.8%) and QA-MRR (82.6%), while GPT-6-sol has the highest QA-Acc (70.6%).

Open-source model rankings also depend on the framework and metric. InternVL-3.5 8B achieves the highest open-source Ans-F1 under Video-RAG (81.7%) and EgoButler (61.4%). Qwen-3.8 27B leads open-source QA-Acc under both frameworks and leads all three metrics under WorldMM. Compared with Qwen-3.5 9B, Qwen-3.8 27B scores higher on all three Video-RAG and WorldMM metrics. Under EgoButler, however, its Ans-F1 falls from 53.0% to 46.0%, and its QA-MRR falls slightly from 61.5% to 61.3%, despite a small QA-Acc increase from 40.7% to 41.1%. These comparisons show that gains in one framework do not necessarily carry over to another.

Task-level results highlight the value of structured retrieval. Figure 6 compares accuracy across six task categories for Video-RAG, EgoButler, and WorldMM. Video-RAG shows relatively balanced performance, while EgoButler varies more across tasks and base models, particularly on timeline reconstruction. WorldMM is strongest on conversational memory and intent recall: with Gemini-3.8-Flash, it reaches 90.7% and 84.4%, respectively, compared with 65.7% on timeline reconstruction and 63.8% on object location. These gaps show that gains in conversational and intent recall do not extend uniformly to temporal and object-location memory.

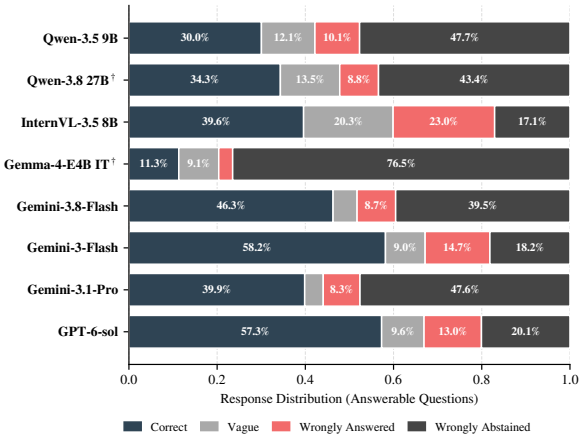


Figure 7: Response reliability on answerable QAs.

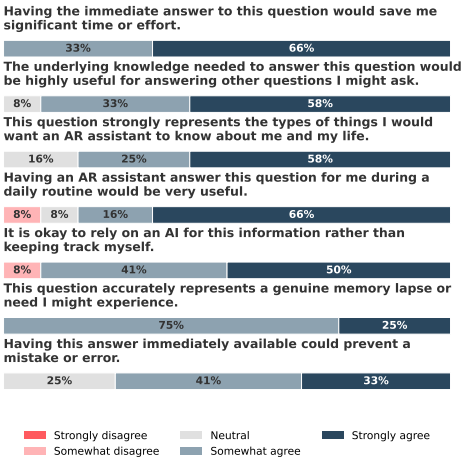


Figure 8: Survey responses for QA quality.

Reliability failures are dominated by excessive abstention. Figure 7 decomposes responses on answerable questions under Video-RAG. For seven of the eight models, wrong abstention is the largest error category: models decline to answer even though supporting evidence exists. For Qwen-3.5 9B, Qwen-3.8 27B, Gemma-4-E4B IT, and Gemini-3.1-Pro, wrong abstentions (43.4–76.5%) exceed correct answers. The two most reliable models, Gemini-3-Flash and GPT-6-sol, answer only 58.2% and 57.3% of answerable questions correctly and wrongly abstain on 18.2% and 20.1%, respectively. Thus, `SuperMemory-VQA` tests whether models know when to answer and when to abstain: they must avoid hallucinating when evidence is absent while also avoiding unnecessary abstention when evidence is available. Appendix 16.5 analyzes these Flash and Pro failure cases qualitatively. In summary, answer reliability depends on abstention behavior as much as on retrieval. Under Video-RAG, the three models that abstain least (InternVL-3.5 8B, Gemini-3-Flash, GPT-6-sol) produce the three highest wrong-answer rates, and no model answers more than 60% of answerable questions correctly. The gap between Ans-F1, QA-Acc, and answer reliability leaves substantial room for systems that couple long-horizon retrieval with grounded, calibrated answering.

5.3 Survey Results

We conducted a survey with the participants to measure perceived question realism, usefulness, and alignment with everyday memory needs. Users reviewed 18 questions from their own recordings and rated seven statements about question quality and utility. As shown in Figure 8, responses were strongly positive, with no disagreement across statements: all participants agreed that questions captured genuine memory lapses, and 82% found the answers useful during daily routines. The survey also suggests that `SuperMemory-VQA` captures reusable personal knowledge rather than isolated QA pairs: 78% agreed that the underlying knowledge would also help answer future questions. This pattern supports the broader motivation of `SuperMemory-VQA`: long-horizon memory should be useful, personal, and grounded, while still requiring careful treatment of user trust and control.

6 Conclusion

In this work, we introduced `SuperMemory-VQA`, an egocentric VQA dataset and benchmark for practical long-horizon memory tasks. It combines multimodal AI glass recordings with temporally grounded questions, evidence annotations, and ordered answer choices that distinguish accurate, vague, incorrect, and unanswerable responses. Our benchmarking results show that current video understanding and retrieval-augmented systems remain unreliable: strong VLMs still struggle with answerability, long temporal gaps, and multi-moment evidence integration. Participant feedback confirms practical daily relevance. Overall, `SuperMemory-VQA` moves evaluation toward grounded, situated memory systems and highlights the need for accurate, hallucination-robust models that answer only when the available evidence is sufficient.

AI USE STATEMENT

In this work, we used generative AI tools as part of dataset construction and experimental evaluation. Specifically, Gemini-3.1-Pro-Preview and Gemini-3-Flash-Preview were used for dense multimodal captioning, candidate question-answer generation, evidence retrieval, criterion-based verification, and iterative refinement of annotations. WhisperX was used to transcribe recorded audio. Generative AI models were also evaluated as benchmark systems. Additionally, generative AI tools assisted with drafting and language editing portions of the manuscript. We did not use generative AI to formulate mathematical claims or proofs; theoretical model development, proof writing, and translation were not applicable to this work.

All AI-generated captions and question-answer annotations underwent human review. Reviewers checked the generated content against the source videos and evidence timelines for factual grounding, temporal causality, answerability, question naturalness, and answer-choice quality, and corrected or rejected erroneous examples. The authors also reviewed the AI-assisted manuscript content and verified the reported experimental results. We take responsibility for the final content of this work, including all text, claims, data annotations, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Data collection was conducted under an IRB-approved protocol, and all participants provided informed consent, including consent for public release of processed recordings. We use pseudonyms, withhold raw audio and participant demographics, redact sensitive transcript content, blur faces and license plates, and remove interactions involving non-consenting individuals. We acknowledge remaining privacy risks and the limited demographic and environmental diversity of the dataset, as discussed in the Appendix.

REPRODUCIBILITY STATEMENT

We release the dataset, annotations, data-processing code, and evaluation scripts. Section 3 and 4 describe dataset construction, baselines, implementation details, and metrics. Appendix 8 provides the collection protocol, hardware specifications, annotation pipeline and schemas, human-review procedure, compute details, evaluation protocol, and additional results needed to reproduce the reported experiments.

REFERENCES

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH*, 2023. doi: 10.21437/interspeech.2023-78. URL https://www.isca-archive.org/interspeech_2023/bain23_interspeech.html.

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

Maria A. Brandimonte, Gilles O. Einstein, and Mark A. McDaniel (eds.). *Prospective Memory: Theory and Applications*. Lawrence Erlbaum Associates, Mahwah, NJ, 1996. doi: 10.1016/s0028-3932(97)80257-6. URL https://linkinghub.elsevier.com/retrieve/pii/S0028393297802576.

Meng Chu, Yicong Li, and Tat-Seng Chua. Graphvideoagent: Enhancing long-form video understanding with entity relation graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4639–4648, 2025. doi: 10.1145/3746027.3755537. URL https://dl.acm.org/doi/10.1145/3746027.3755537.

Neal J. Cohen and Larry R. Squire. Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, 210(4466):207–210, 1980. doi: 10.1126/science.7414331. URL https://doi.org/10.1126/science.7414331.

Neal J Cohen, Russell A Poldrack, and Howard Eichenbaum. Memory for items and memory for relations in the procedural/declarative memory framework. *Memory*, 5(1-2):131–178, 1997. doi: 10.1080/741941149. URL http://www.tandfonline.com/doi/abs/10.1080/741941149.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, September 2018. doi: 10.1007/978-3-030-01225-0_44. URL https://link.springer.com/chapter/10.1007/978-3-030-01225-0_44.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

Howard Eichenbaum. Declarative memory: Insights from cognitive neurobiology. *Annual review of Psychology*, 48(1):547–572, 1997. doi: 10.1146/annurev.psych.48.1.547. URL https://www.annualreviews.org/doi/10.1146/annurev.psych.48.1.547.

Gilles O. Einstein and Mark A. McDaniel. Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):717–726, 1990. doi: 10.1037/0278-7393.16.4.717. URL https://doi.org/10.1037/0278-7393.16.4.717.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. doi: 10.48550/arXiv.2308.13561. URL https://arxiv.org/abs/2308.13561.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, pp. 75–92. Springer, 2024. doi: 10.1007/978-3-031-72670-5_5. URL https://link.springer.com/10.1007/978-3-031-72670-5_5.

Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 314–327. Springer, 2012. doi: 10.1007/978-3-642-33718-5_23. URL http://link.springer.com/10.1007/978-3-642-33718-5_23.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025. doi: 10.1109/cvpr52734.2025.02245. URL https://ieeexplore.ieee.org/document/11093290/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. doi: 10.48550/arXiv.2312.10997. URL https://arxiv.org/abs/2312.10997.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022. doi: 10.1109/cvpr52688.2022.01842. URL https://ieeexplore.ieee.org/document/9879279/.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, pp. 19383–19400, 2024. doi: 10.1007/s11263-025-02557-6. URL https://link.springer.com/10.1007/s11263-025-02557-6.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 292–299. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-4340. URL https://doi.org/10.3115/v1/W14-4340.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, et al. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24169–24179, 2025. doi: 10.1109/cvpr52734.2025.02251. URL https://ieeexplore.ieee.org/document/11093318/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021. doi: 10.48550/arXiv.2112.09118. URL https://arxiv.org/abs/2112.09118.

JaidedAI. EasyOCR: Ready-to-use OCR with 80+ supported languages. https://github.com/JaidedAI/EasyOCR, 2020. URL https://github.com/JaidedAI/EasyOCR.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019. doi: 10.1109/tbdata.2019.2921572. URL https://ieeexplore.ieee.org/document/8733051/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397.

Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 287–295, 2015. doi: 10.1109/cvpr.2015.7298625. URL http://ieeexplore.ieee.org/document/7298625/.

Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 619–635, 2018. doi: 10.1007/978-3-030-01228-1_38. URL https://link.springer.com/10.1007/978-3-030-01228-1_38.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00638/119630/Lost-in-the-Middle-How-Language-Models-Use-Long.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024. doi: 10.48550/arXiv.2411.13093. URL https://arxiv.org/abs/2411.13093.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. doi: 10.48550/arXiv.2402.13349. URL https://arxiv.org/abs/2402.13349.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024. doi: 10.1007/978-3-031-72691-0_25. URL https://link.springer.com/10.1007/978-3-031-72691-0_25.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, volume 36, pp. 46212–46244, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

Karttikeya Mangalam, Jitendra Malik, et al. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *arXiv preprint*, 2023b. URL https://arxiv.org/abs/2308.09126.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. Egoblur: Responsible innovation in aria, 2023. URL https://arxiv.org/abs/2308.13093.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018. doi: 10.18653/v1/p18-2124. URL https://aclanthology.org/P18-2124/.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M de Melo, Benjamin Van Durme, and Rama Chellappa. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19691–19701, 2025. doi: 10.1109/cvpr52734.2025.01834. URL https://ieeexplore.ieee.org/document/11094542/.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. doi: 10.48550/arXiv.2502.01549. URL https://arxiv.org/abs/2502.01549.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13193–13203, 2024. doi: 10.1109/cvpr52733.2024.01253. URL https://ieeexplore.ieee.org/document/10658173/.

Larry R. Squire. Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82(3):171–177, 2004. doi: 10.1016/j.nlm.2004.06.005. URL https://doi.org/10.1016/j.nlm.2004.06.005.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=uyTL5Bvosj.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. doi: 10.48550/arXiv.2403.05530. URL https://arxiv.org/abs/2403.05530.

Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory*, pp. 381–403. Academic Press, 1972. doi: 10.4135/9781446212967.n15. URL https://sk.sagepub.com/books/cognitive-psychology/n15.xml.

Endel Tulving. Episodic memory: From mind to brain. *Annual Review of Psychology*, 53(1):1–25, 2002. doi: 10.1146/annurev.psych.53.100901.135114. URL https://doi.org/10.1146/annurev.psych.53.100901.135114.

David Wan, Han Wang, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Clamr: Contextualized late-interaction for multimodal content retrieval. *arXiv preprint arXiv:2506.06144*, 2025. doi: 10.48550/arXiv.2506.06144. URL https://arxiv.org/abs/2506.06144.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, pp. 58–76. Springer, 2024. doi: 10.1007/978-3-031-72989-8_4. URL https://link.springer.com/10.1007/978-3-031-72989-8_4.

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, 2023a. doi: 10.1109/iccv51070.2023.01854. URL https://ieeexplore.ieee.org/document/10377735/.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *CoRR*, abs/2307.09702, 2023. doi: 10.48550/arXiv.2307.09702. URL https://arxiv.org/abs/2307.09702.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 808–819. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1078. URL https://doi.org/10.18653/v1/P19-1078.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yong Liu, Xiangtai Li, Dacheng Tao, et al. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *NeurIPS*, 2025. doi: 10.48550/arXiv.2506.13589. URL https://openreview.net/forum?id=FDAI0PY9Qp.

Charig Yang, Samiul Alam, Shakhrul Iman Siam, Michael J Proulx, Lambert Mathias, Kiran Somasundaram, Luis Pesqueira, James Fort, Sheroze Sheriffdeen, Omkar Parkhi, Yuheng Ren, Mi Zhang, Yuning Chai, Richard Newcombe, and Hyo Jin Kim. Reading recognition in the wild. In *Advances in Neural Information Processing Systems*, 2025a. URL https://nips.cc/virtual/2025/poster/117940.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *CVPR*, pp. 28885–28900, 2025b. doi: 10.1109/cvpr52734.2025.02690. URL https://ieeexplore.ieee.org/document/11095171/.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

7 APPENDIX

8 DATA COLLECTION SUPPLEMENTARY INFORMATION

This section supplements the dataset overview and collection procedure in §3.1 and §3.3.

8.1 HARDWARE AND MODALITIES.

Participants wore Gen 1 Meta Aria Glasses to capture high-frequency, egocentric multimodal data. The collected raw data streams included one 1408×1408 RGB video stream collected at 30fps, two 640×480 Grayscale video stream for SLAM collected at 30fps, a 320×240 eye-tracking camera array operating at 60fps to monitor gaze behavior, and 7 channels of high-fidelity audio sampled at 48kHz. Additionally, the device logged motion and orientation data via two Inertial Measurement Units (IMUs) operating at 1000Hz and 800Hz, a Magnetometer at 10Hz, and a Barometer at 50Hz. We also computed open and closed loop trajectories, as well as 3D point clouds of the entire space from SLAM and IMU motion data. Eye tracking data were also obtained from processing the raw eye tracking camera streams.

8.2 PROTOCOL.

Data collection was conducted under a protocol approved by the Institutional Review Board (IRB). The released dataset described in this paper contains recordings from 10 participants recruited from the general population and university settings. The IRB protocol frames the study as an evaluation of the memory capabilities of an AI-driven wearable system, "Supermemory," that integrates multimodal sensor data from Meta Aria Glasses with long-term memory retrieval. Unlike studies that focus on perception within a fixed timespan, the protocol targets dynamic, non-linear memory formation across everyday activities over longer time spans. Using synchronized high-frequency eye-tracking, audio, IMU, RGB, and SLAM video data, the study analyzes how contextual and temporal cues influence recall, intent tracking, and memory construction. The protocol allowed either single-subject recordings or multiple participants working together and conversing while wearing the same type of AR glasses.

Potential participants were identified through recruitment materials and contacted the study team by email if interested. The study team confirmed eligibility before scheduling sessions, shared the types of tasks participants would perform, discussed whether participants were comfortable with outdoor tasks or tasks involving other participants, and noted preferences about study locations and timing. During screening, participants were also told how the data would be processed and that residual identifiability could remain even after processing.

Participants completed an informed consent process before any study procedures using a legally valid electronic signature workflow. After eligibility was confirmed and a session was scheduled, the study team emailed a DocuSign link to the consent form. Participants were encouraged to review the consent form in advance and could ask questions by email, phone, or in person before signing. At check-in, the study team confirmed the participant's identity and completion of the correct consent form; a copy of the signed consent was automatically provided through DocuSign. The consent process emphasized that participation was voluntary, that participants could decline or discontinue at any time without penalty, and that compensation was not contingent on completing tasks beyond the prorated time completed. Participants were also informed that participation required consent for face-blurred study data to be shared publicly through the Hugging Face Hub; individuals unwilling to allow public release of their processed data were asked not to participate.

Data was collected in a short-term rental (Airbnb) to simulate a home environment, with both indoor and outdoor segments included under the IRB and non-participant privacy protocol. Each recording session had at least one member of the lab participating in it to guide the workflow with one to three other external participants. All participants calibrated their glasses before starting to record. Participants were then asked to explore the environment and take quick notes of where everything was. After that, they were explained what to do based on a predefined script, but were encouraged to loosely follow and keep all interactions natural. Scripts included cooking instructions for preparing a recipe, manuals for playing a board game, or assembling a puzzle. Participants read or were orally

dictated the specific instructions by the lab member and were asked to follow them. Participants planned among themselves how to implement the script and collaboratively finish it. To prevent leaks of sensitive information, participants were assigned code names and used those names instead of their real names. After 30 to 45 minutes, participants took breaks and then changed glasses before continuing.

The protocol allowed follow-up sessions using the same general procedure with different scripts, either to collect additional task permutations or to replace unusable data. Follow-up participation was voluntary, and participants could decline further sessions. Participants could incur parking or transportation costs, which were not reimbursed. Incentives were provided as physical Amazon gift cards and were prorated when participation was incomplete. Researchers who collected protocol-testing data as study subjects did not receive incentives.

The protocol describes the study as minimal risk and non-invasive. The main physical risks were slight fatigue from wearing the glasses or performing scripted tasks, mitigated by breaks between scripts. The Aria glasses include onboard temperature sensors and shutoff behavior for overheating, and attendants monitored glass temperature during sessions. The protocol identifies no direct benefit to individual participants, but describes societal benefits from enabling future AR systems that support long-term episodic recall and situational awareness and from providing an open dataset to accelerate research in human-centered AI. Primary endpoints include recall accuracy measured by automated vision-language models against human-verified ground truth from the scripted scenarios.

The full participant-facing protocol materials, including task scripts, consent language, and supplementary study materials, are included with the supplemental submission.

8.3 Privacy and Anonymization Steps.

To minimize the publication of identifiable and sensitive information, we publish the RGB video, processed eye gaze data, trajectory data, SLAM point-cloud data, and IMU data only. We do not release the raw audio data, but instead provide transcribed text from audio using the WhisperX (Bain et al., 2023) model and manually remove any private information accidentally divulged by participants. We also blur faces and license plates from the RGB video data using the EgoBlur (Raina et al., 2023) model. Outdoor captures of non-participants from passing glances were blurred. However, any direct interaction, such as talking with non-participants, was manually removed from the recordings.

The IRB protocol used a multi-layer privacy plan for incidental capture of non-participants. Public-facing activities were scheduled, when possible, during off-peak times and in low-traffic areas, and the study avoided areas with heightened privacy expectations. For indoor recordings, any segment containing facial data from a non-consenting person was removed from the release. For necessary outdoor public recordings, faces were blurred automatically and then manually verified so missed faces could be removed or redacted. The study team did not actively interact with non-participants; any direct interaction with an outside person was removed from the released dataset.

We used a multi-stage de-identification and quality-control process. Automated tools, including EgoBlur and multimodal LLMs, were used to detect and mask faces and to flag frames containing potentially identifying text such as IDs, name tags, license plates, or personal account information on screens. We then manually reviewed outputs to identify missed faces, unwanted segments, and sensitive text. Only processed data with blurred faces and transcribed, redacted audio is shared broadly.

9 Task Description Supplementary Information

This section expands the task taxonomy introduced in §3 and compared against prior datasets in §3.2.

`SuperMemory-VQA` dataset categorizes question-answer pairs into six distinct functional tasks. Each task mimics a real-world application of human memory augmentation. Below, we describe these tasks.

Object & Location Memory. This task requires the system to identify the last known position of a specific object or track its movement across different time spans and locations, maintaining object permanence. This task is based on the concept of episodic memory (Tulving, 1972) and has been

previously used in egocentric vision (Grauman et al., 2022). It evaluates an AI agent's ability to index a wearer's data and locate an object based on its last observed time. This effectively augments human memory on demand, answering queries such as, "I cannot find the black scissors. Where did I leave it last?" or "I forgot where I put my keys. Where did I put it?".

Conversational Memory. Systems must also recall specific facts from multi-topic chats, including tracking commitments, deferred answers, and mid-conversation corrections. Conversational Memory, also known as Dialogue State Tracking (DST), is a formal component within task-oriented conversational AI (Henderson et al., 2014; Wu et al., 2019). Robust DST is critical for maintaining awareness across exchanges without losing context due to natural conversational shifts. Examples of this task include queries like, "Where did I say we should meet after this?" or "What did B tell me to get after I finished working on this?".

Visual Scene Recall. This task evaluates the ability to recall specific visual details from past environments, such as text on whiteboards, instructions in a manual, or displayed on a screen. While rooted in episodic memory (Tulving, 1972), it concurrently evaluates the system's *semantic memory* (Tulving, 1972; Squire, 2004). To successfully manage dense information retrieval, the AI must go beyond visual recall to recognize objects, read text, and link these elements to general world knowledge and factual meaning independent of the temporal context. Evaluated examples include, "Was I supposed to add the spices now or after frying?" or "I forgot how many cups of rice I put in the pot. How much did I put in?".

In-Context Retrieval. This task requires the system to perform multi-hop reasoning by chaining multiple disjoint facts retrieved from the user's history. It evaluates the system's capacity for *relational memory* (Eichenbaum, 1997; Cohen et al., 1997)—the cognitive ability to represent and navigate associations between independent elements of an experience. Unlike simple retrieval, in-context retrieval requires the AI to identify a primary fact (e.g., a mentioned time or location) and use it as a prerequisite context to evaluate or retrieve a secondary piece of information (e.g., the status of an ongoing task). This task models complex real-world situational awareness and planning where memory serves as a substrate for logical deduction. Example queries include, "Given the meeting time B mentioned, do I have time to finish my presentation?" or "My bus leaves in 20 minutes. Will the dryer finish up in time?".

Timeline Reconstruction. The agent is required to sequence events chronologically, such as listing all locations visited during a multi-location errand in the correct order. Timeline reconstruction directly evaluates the temporal aspect of episodic memory (Tulving, 2002) through the temporal localization and chronological sequencing of disjointed events across a longer horizon. Furthermore, when applied to structured activities, this task models the tracking of *procedural memory* (Cohen & Squire, 1980; Squire, 2004)—the implicit knowledge of how to perform tasks and specific action sequences. By reconstructing timelines, the agent demonstrates an understanding of the step-by-step procedures required to complete an activity. For instance, the system might be asked, "Did I miss any ingredients when preparing the marinade?" or "Does anyone own this property on the Monopoly board?".

Intent Recall. Systems must retrieve information based on explicit, verbal reminders or evaluate the passive recall of goals that were implied but ultimately not completed by the user. This aligns directly with *prospective memory*, the cognitive process that allows individuals to remember to perform intended actions in the future (Einstein & McDaniel, 1990; Brandimonte et al., 1996). Prospective memory acts as a bridge between intentions and actions, requiring continuous monitoring to execute tasks at the appropriate time. Example queries include explicit intent like, "What did I intend to do after I finished this?" or implicit reminders like "Washer cycle is done. Remind user to put the clothes into the dryer." Reminders can be tied to a specific point in time or be triggered based on location or interaction with a person.

To validate that our task categories reflect distinct memory behaviors, we asked participants to answer a representative sample question from each task and describe the thought process they would use if answering from memory. The resulting mapping between task categories and recalled reasoning strategies is summarized in Figure 5. The flows show that the taxonomy is not only based on surface question wording: different tasks elicit different memory operations, such as locating an object, reconstructing temporal order, checking prior intent, or integrating evidence across events.

This supports our categorization of SuperMemory-VQA tasks as distinct forms of long-horizon egocentric memory.

10 LIMITATIONS

This section complements the benchmark-scope discussion in §3.1.

SuperMemory-VQA is an initial benchmark for long-horizon egocentric memory rather than an exhaustive representation of all daily-life settings. The current release contains 52.9 hours of recordings from 10 participants, which is large enough to expose important retrieval, grounding, and answerability failures but still limited relative to the diversity of homes, cultures, languages, activities, and accessibility needs that future AR memory assistants will encounter. The dataset also relies on loosely scripted indoor and outdoor activities centered around a simulated home setting, so it may not capture the full unpredictability, privacy sensitivity, and behavioral diversity of unconstrained daily life.

At the same time, the benchmark is designed as a building block. The annotation pipeline separates dense captioning, structured ledger construction, agentic QA generation, verification, and human review. This modular design can be reused for future data collection efforts and scaled to additional participants, longer recordings, new environments, and specialized use cases. We view SuperMemory-VQA as a foundation for measuring current system failures and for guiding larger future benchmarks for grounded AR memory with explicit evidence-sufficiency checks.

11 EVALUATION REPRODUCIBILITY AND COMPUTE

This section supplements the experimental setup in §4 and the main results in §5.1.

We evaluate all systems using the metrics reported in the main paper: answerability F1, QA accuracy, and QA mean reciprocal rank. Open-source model evaluations were run on a server with $4\times$A100 GPUs. Gemini-family closed-source models were accessed through the Google Cloud Platform API, and OpenAI-family models were accessed through Azure OpenAI APIs. The released code repository contains the evaluation scripts needed to reproduce the reported baseline comparisons.

12 ANNOTATION PIPELINE SUPPLEMENTARY INFORMATION.

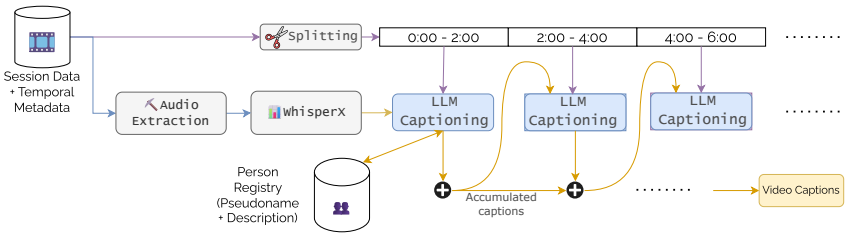
This section expands the annotation pipeline summarized in §3.4. To scale the generation of question-answer pairs in SuperMemory-VQA, we developed an automated, agentic question-answer generation pipeline. This pipeline leverages multiple specialized LLM agents and is structured into two phases. Each phase goes through human review to maintain high data fidelity.

12.1 PHASE 1: DENSE VIDEO CAPTIONING.

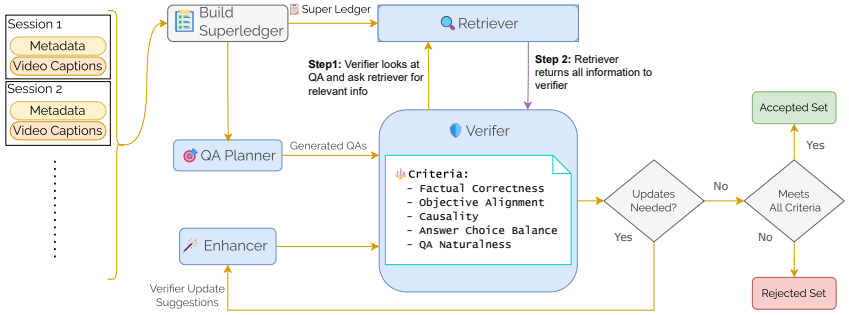
The pipeline initializes with egocentric Session Videos and Temporal Metadata (video recording time and date, duration, etc). Due to the context limits of LLMs, these videos are first split into shorter chunks. Concurrently, an *Audio Extraction* process is performed, with the audio of all sessions for a subject combined together and then passed through *whisperX* for transcription. An *LLM Captioning* agent then processes each video chunk along with the transcription, a *Person Registry* (containing pseudonames and descriptions of the individuals as they are seen in the recordings), and previously accumulated captions. This agent extracts dense descriptions of visual actions, detected objects, auditory events, and conversation summaries. These discrete captions are subsequently temporally aggregated, producing consolidated Video Captions for the session. New persons are added to the registry when they are seen. The first phase is illustrated in Figure 9a. To ensure the text accurately reflects the raw video, this intermediate output undergoes an initial stage of *Human Review*.

12.2 PHASE 2: AGENTIC QA GENERATION.

Phase 2 consists of 4 agents: *QA PLanner*, *Verifier*, *Retriever*, and *Enhancer*. It is illustrated in the diagram in Figure 9b. It begins with a *Build Superledger* process, which aggregates the Metadata and



(a) Phase 1: Dense Video Captioning



(b) Phase 2: Agentic QA Generation

Figure 9: Illustration of the question answer generation pipeline.

Video Captions from all sessions into a unified "Super Ledger". A *QA Planner* agent reads this Super Ledger to propose diverse question-answer pairs designed to target the dataset's core dimensions and tasks (Wang et al., 2023b). We employ a rationale-first schema design principle during question answer generation (Willard & Louf, 2023). We enforce *instance-level chain-of-thought* by embedding reasoning fields before generating an annotation with a question and answer data (Xiao et al., 2024). This localizes the thought process of the agent to the local context.

Note that the QA Planner only uses the text of the Superledger for this. So, to ensure that the generated pairs are grounded in the source media, the pipeline employs a closed-loop verification system. For each generated annotation, a *Verifier* agent asks a *Retriever* for relevant information. The Retriever searches the Super Ledger and returns all relevant data back to the Verifier. The Verifier then evaluates the proposed annotation against a strict set of criteria. Factual correctness ensures that the information is grounded in the data. Objective alignment checks if the question answer pair is relevant to the objective of `SuperMemory-VQA`, specifically if it adheres to the predefined tasks and if the temporal gap is significant enough to justify a possible memory lapse. Causality criteria ask the verifier to make sure the question is answerable using only previously recorded data. Answer Choice Balance asks the Verifier to ensure the answer cannot be simply guessed by looking at only the question and answer choices. Finally, QA Naturalness criteria make sure that the question is actually a practical and naturally worded question that a user may ask. The Verifier also follows the same principle as the QA planner and generates an individual criterion score reason before generating the actual score (Zheng et al., 2023). Based on this evaluation, the Verifier determines if updates are needed. If updates are required, the pair is routed to an *Enhancer* agent, which updates question-answer pairs according to the verifier's suggestions and feeds them back to the Verifier in an iterative loop. If the verifier gives no suggestions and the annotation is marked as incorrect, the pair is considered unusable and routed to the *Rejected set*. If the annotation is marked correct and has no further suggestions, it is considered approved. Approved pairs are routed to the *Accepted Set*. Finally, the pairs in the Accepted Set undergo a final *Human Review* phase to yield the definitive benchmark dataset.

The QA Planner and Verifier agents use the gemini-3.1-pro-preview model, whereas the LLM Captioning Agent, Video Retriever, and Enhancer agents use gemini-3-flash-preview.

12.3 PIPELINE COST ANALYSIS.

We estimate the variable API cost of the annotation pipeline from duration-based media tokenization rather than frame counts. This distinction matters for long egocentric recordings: in the current Gemini API accounting, video and audio inputs are converted at fixed rates of 263 and 32 tokens per second, respectively, for a combined media density of 295 tokens per second for video with audio.[2] Pricing is computed using the May 2026 Gemini Developer API standard paid tier: gemini-3-flash-preview charges $0.50/M input tokens for text/image/video, $1.00/M input tokens for audio, and $3.00/M output tokens; gemini-3.1-pro-preview charges $2.00/M input tokens and $12.00/M output tokens for prompts up to 200k tokens, with higher prices for larger prompts.[3] Unless otherwise stated, the estimates below use online standard pricing, assume no Batch/Flex discounts, and exclude human review, local preprocessing, storage, taxes, and non-LLM compute.

Cost model. Stage 1 splits each hour into thirty 120s chunks and runs captioning on Flash. Each chunk contributes $120 \times (263 + 32) = 35.4$k media tokens plus approximately 5k prompt/context tokens. Stage 2 includes one Pro QA-generation pass over the Super Ledger, followed by a closed-loop verification handshake: a text-only verifier request, a Flash ledger-retrieval request, one Pro verifier pass over 10 evidence clips totaling 300s, and one Flash enhancer pass over the same evidence budget. We assume three verification loops per QA, 90% cache reuse for the session-level Super Ledger, and 15% reuse for recurring evidence clips. Output lengths are estimated at 2k tokens per caption chunk, 1.5k tokens for QA generation, and 0.5–1.5k tokens per verification-loop agent call.

Table 3: Unit cost assumptions for the SuperMemory annotation pipeline. Costs use current Gemini standard-tier pricing and duration-based video/audio tokenization.

Unit	Dominant inputs	Input tokens	Output tokens	Est. cost
Stage 1: Dense video captioning				
1 hour, Flash captioner	30 chunks; 946.8k video tokens, 115.2k audio tokens, 150k text/context tokens	1.212M	60k	$0.84
Stage 2: QA generation, per QA				
QA generation	Super Ledger text at $\sim$ 15k tokens/hour; 90% cached, gemini-3.1-pro-preview	$15T$+1k	1.5k	$0.0200 + 0.0057T$
Stage 2: Closed-loop verification, per QA per loop				
Verifier setup	QA pair and rubric text on gemini-3.1-pro-preview	1.0k	0.5k	$0.008
Retriever	Super Ledger text at $\sim$ 15k tokens/hour; 90% cached, gemini-3-flash-preview	$15T$+1k	1.0k	$0.0035 + 0.001425T$
Verifier evidence pass	10 clips totaling 300s plus prompt; 15% cached, gemini-3.1-pro-preview	93.5k	1.5k	$0.180
Enhancer evidence pass	Same evidence budget on gemini-3-flash-preview	93.5k	1.5k	$0.049
Total loop	Pro verifier with Flash retriever/enhancer	195.5+$15T$k	4.5k	**$0.2407 +$0.001425T**

Note. T denotes total session duration in hours. For T=1, QA generation costs approximately $0.026 and one closed-loop verification iteration costs approximately $0.242; at three loops, generation plus verification is $0.752 per QA. At T=50, ledger context dominates the text-only generation/retrieval steps, raising generation plus three verification loops to approximately $1.241 per QA.

The dominant cost driver is therefore not Stage 1 captioning but repeated multimodal evidence verification and Pro QA generation over a long Super Ledger. For the default configuration, captioning

[2] https://ai.google.dev/gemini-api/docs/tokens

[3] https://ai.google.dev/gemini-api/docs/pricing

Table 4: Worst case scale projection for the default model mix: Flash captioning, retrieval, and enhancement; Pro QA generation and verification; and three verification loops per QA.

Hours	QA density	QAs	Total tokens	Default cost
Small-to-medium annotation runs				
1	0.5/min	30	21.1M	$23.40
1	1.0/min	60	41.0M	$45.97
1	2.0/min	120	80.8M	$91.09
2	0.5/min	60	45.9M	$47.41
2	1.0/min	120	89.2M	$93.13
2	2.0/min	240	175.9M	$184.57
10	0.5/min	300	373.5M	$260.98
10	1.0/min	600	734.2M	$513.52
10	2.0/min	1,200	1.456B	$1,018.60
50	1.67/min	5,000	18.076B	**$6,246.18**

The 50 hour, 5,000 QA extrapolation is intentionally conservative for text context: it sends a 15k-token/hour Super Ledger prefix during Pro QA generation and each Flash retrieval loop, with cache-aware pricing. If the planner or retriever first narrows the ledger with an index or symbolic search before calling Gemini, the Stage 2 text cost drops; if evidence bundles exceed 300s or Pro requests cross the 200k-token pricing threshold, costs rise. Note, the above data is for the worst case when all QA pairs are resent for enhancement. From our experience, this number falls by 30-50% after every iteration. **The approximate cost for our data was around $3900**

50 hours costs approximately $42, whereas Pro QA generation plus three-loop verification for 5,000 QAs costs approximately $6.20k under the cache-aware ledger model. This suggests that future scaling should prioritize (i) evidence-clip pruning before verifier calls, (ii) reusing cached high-value evidence clips across related QAs, and (iii) indexing the Super Ledger before QA generation and retrieval.

12.4 Verification Criteria

To guarantee that the SuperMemory-VQA benchmark meets the highest standard of scientific rigor, we employ a multi-stage validation pipeline consisting of agentic scoring, deterministic temporal filtering, structural schema verification, and visual spatial-grounding validation.

1. **Agentic Quality Scoring (Stage 2 Verifier Agent)**: The candidate QA pairs are first evaluated across three continuous dimensions ($\in [0.0, 1.0]$) by a specialized Verifier Agent:
   - *Factual Correctness:* Ensures complete alignment between the answer text and visual/audio evidence.
   - *Objective Relevance:* Evaluates question logicality, requiring a minimum temporal gap between evidence and query (> 10 minutes) and multi-clip reasoning.
   - *Causal Answerability:* Verifies that the question can be resolved using only the evidence available up to the question's execution timestamp.

   A candidate pair is rejected if any score falls below the threshold $\tau = 0.6$.
2. **Deterministic Causal Filtering and Temporal Grounding**: Beyond agentic intuition, the pipeline programmatically enforces strict physical causality:
   - *Causal Evidence Pruning:* A deterministic filter calculates the absolute timeline offsets of all video chunks. Any answer evidence whose start timestamp t_{evidence} is greater than the question start timestamp t_{question} is programmatically pruned from the final dataset.
   - *Anchor Validity:* All timestamps (MM:SS) must strictly fall within the boundaries of the associated video segments, preventing temporal hallucinations or drift.
3. **Structural and Semantic Integrity Verification**: Every verified pair is validated against a strict Pydantic-enforced schema to guarantee downstream compatibility:

- *Multiple Choice Divergence:* Every QA pair must contain exactly three choices. For answerable queries (is_answerable = True), the choices must consist of exactly one 'correct', one 'vague' (technically correct but general/ambiguous), and one 'incorrect' distractor. For unanswerable queries (is_answerable = False), all three choices must be classified as 'incorrect' (plausible-sounding but hallucinated distractors) to test whether models can recognize information boundaries without hallucinating. Each choice must also contain a rigorous 'explanation' field justifying its category.

- *Coherent Room & Modality Mapping:* Both questions and answers are validated for spatial and sensory consistency by requiring a designated room/location tag and a combination of active sensor modalities (e.g., *Video, Audio, Gaze, Trajectory, Depth, OCR*).

- *Task-Category Alignment:* Every pair is strictly classified under one of our six defined task categories (e.g., *object_location_memory, timeline_reconstruction*), and checked for alignment with that category's specific reasoning patterns.

Only QA pairs passing all agentic, programmatic and structural criteria are kept in the final output.

12.5 Detailed Input/Output Format

Question-Answer pairs in `SuperMemory-VQA` are multi-modal, i.e., both question and answers may require input from multiple modalities to comprehend. In addition to the text, questions reference a time span, indicating where the query may realistically be asked. Each question also has metadata like location details and relevant sensor modalities needed to understand the question, e.g., Video, Audio, Gaze.

Answers in `SuperMemory-VQA` are naturally phrased multiple-choice sentences. Each question has four ordered choices:

Correct	An accurate answer that directly addresses the query.
Vague	A technically plausible answer that is ambiguous, underspecified, or less useful.
Wrong	An answer that is inconsistent with the evidence.
N/A	An abstention option for queries that lack sufficient grounding in the available evidence.

Each answer is grounded in video evidence, which can span multiple sessions to evaluate long-term memory and recall. Each evidence item includes a text description, referenced session, and video IDs, video time span, location, and relevant sensor modalities. The evidence for an answer always occurs before the question starts. The entire schema is shown in 12.5.

Final Verified QA Pair

```
OutputQAPairFormat
├─balance_reasoning: String
├─question: Object
│ ├─text: String first-person query
│ ├─question_reasoning: String temporal justification
│ ├─room: String ledger environment
│ ├─video_id: String primary video
│ ├─modalities[]: List[Modality]
│ └─time_spans[]: List[Object]
│   ├─video_id: String
│   ├─start_time: String MM:SS
│   └─end_time: String MM:SS
└─answer: Object
  ├─text: String
  ├─is_answerable: Boolean
  ├─evidence_list[]:
  │ List[Object]
  │ ├─reason: String ledger-grounded evidence description
  │ ├─room: String
  │ ├─video_id: String
  │ ├─time_span: Object
  │ │ ├─start_time: String MM:SS
  │ │ └─end_time: String MM:SS
  │ └─modalities[]: List[Modality]
  └─answer_choices[]:
    List[Object]
    ├─explanation: String
    ├─choice_ltype: ChoiceType
    └─text: String
```

```
OutputQAPairFormat
└─metadata: Object
  ├─task: MemoryTask
  ├─task_reasoning: String
  ├─primary_video_id: String
  └─verification_score: Object
    ├─factual_correctness_score: Float
    ├─factual_correctness_reasoning: String
    ├─objective_correctness_score: Float
    ├─objective_correctness_reasoning: String
    ├─causal_answerability_score: Float
    ├─causal_answerability_reasoning: String
    ├─naturalness_score: Float
    ├─naturalness_reasoning: String
    ├─is_correct: Boolean
    ├─is_guessable: Boolean
    ├─guessability_justification: String
    ├─contains_pii: Boolean
    ├─privacy_audit_reasoning: String
    ├─is_salvageable: Boolean
    ├─suggestions[]: List[String]
    └─suggested_chunks[]:
      List[Object]
      ├─video_id: String
      ├─start_time: String MM:SS
      ├─end_time: String MM:SS
      └─relevance_reason: String
```

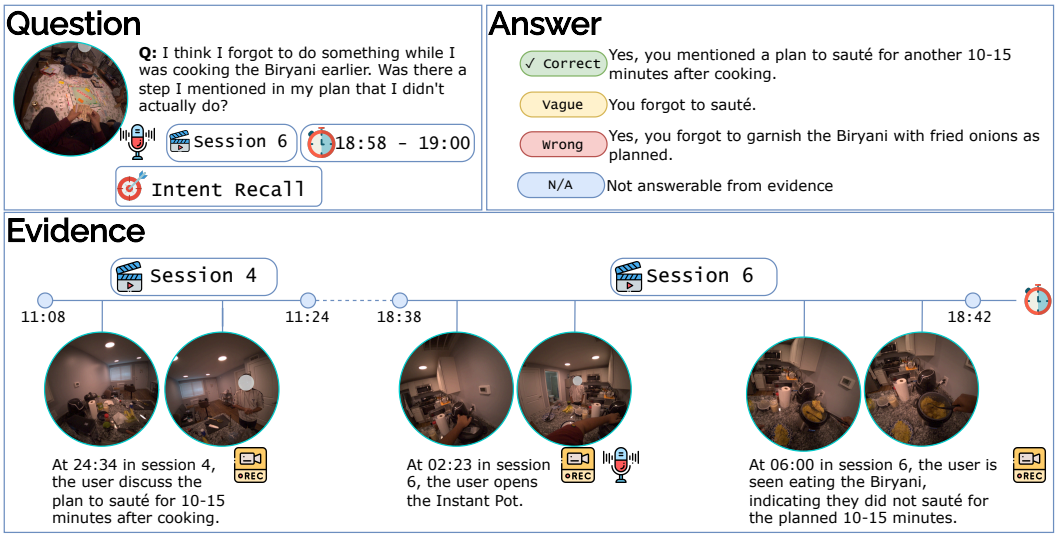


Figure 10: Question Answer Annotation in SuperMemory-VQA

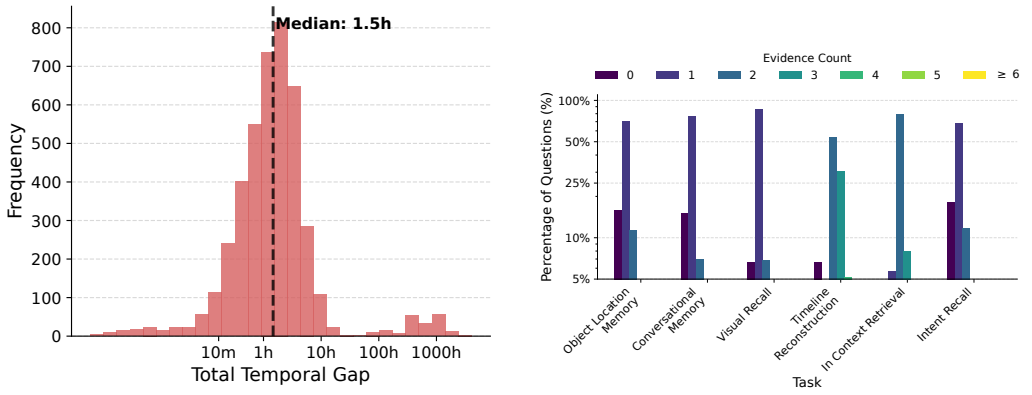
13 Dataset Statistics Supplementary Information

This section expands the dataset-statistics summary in §3.5.

Evidence Complexity: The amount of context a model needs depends heavily on the task (Figure 11b). Tasks like finding a lost object usually just require grounding in a single segment in time or space. However, complex tasks like Timeline Reconstruction force models to piece together clues across multiple separate clips.

Temporal Reasoning Horizons: `SuperMemory-VQA` benchmark is also designed so that answering a question requires searching long stretches of time. We mapped out the "temporal gap"—the time between a user asking a question and when the actual evidence required to answer it was recorded (Figure 11a). While our dataset includes extreme cases where models may have to recall events recorded over a week ago, most questions fall in the 1 to 2 hour range.

Evidence Duration: In addition to the number of evidence clips, we measure how much total video time must be inspected to answer each question (Figure 12a). This statistic captures a complementary source of difficulty: some questions can be grounded by a short visual moment, while others require tracking longer activities or aggregating observations over extended periods. Breaking evidence duration down by task category further shows which categories require sustained temporal grounding rather than isolated retrieval (Figure 12b).



(a) Temporal Gap Distribution

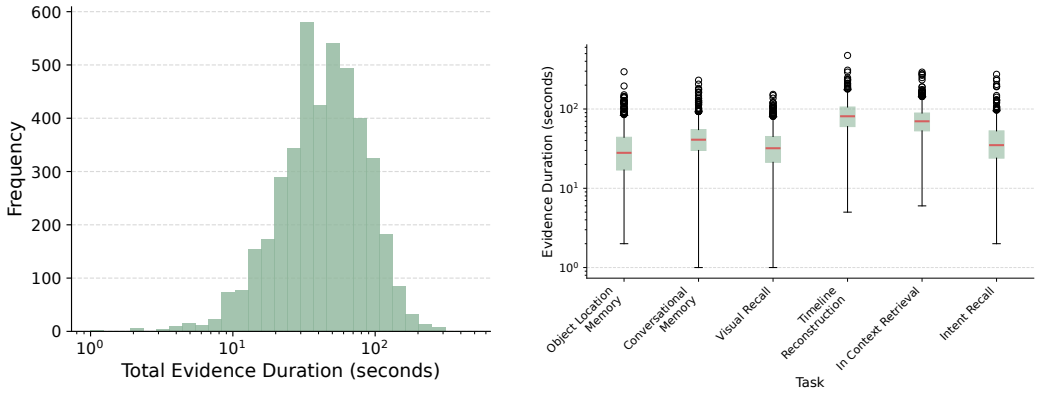
(b) Evidence Count by Task

Figure 11: Overview of dataset statistics (a) temporal gap (in hours) between the timestamp of the visual evidence and the user's query, and (b) the number of distinct evidence clips required to answer a question.

13.1 Human Review.

Human review is the final quality-control step after each automated annotation phase. The same review platform supports two reviewer workflows (Figures 13 and 14). After Phase 1, reviewers use the caption-review interface to inspect the dense video captions produced for each session. The interface loads a processed video, the corresponding caption file, a local timeline window, and the caption chunks generated for that interval. Reviewers can mark a caption file as accepted, rejected, or pending, add new captions, and edit, split, link, or delete individual chunks. In edit mode, each caption exposes structured fields for activities, objects, environment, people, audio transcripts, and timestamps. This stage checks that the captions preserve the visual activity, objects, people, environment, audio events, and text observations needed for downstream question generation.

After Phase 2, reviewers use the QA-review interface to inspect generated question-answer annotations against the source video and evidence timeline. The platform displays accepted and rejected annotation files, the video source, dense temporal markers for questions and supporting evidence, and a side panel for the current annotation. Reviewers can navigate among annotations, review task labels and verification scores, inspect answer options, add or delete annotations, and open raw model



(a) Evidence Duration Distribution

(b) Evidence Duration by Task

Figure 12: Additional dataset statistics showing (a) the total duration of evidence clips required to answer each question, and (b) the distribution of evidence duration across task categories.

responses or model reasoning when debugging failures. In edit mode, reviewers can update the task category, question, reasoning, question time span, answer text, answer choices, evidence spans, modalities, bounding boxes, and human-review status before saving. This workflow lets reviewers verify grounding, factual correctness, answerability, and option quality before examples are included in the final benchmark.

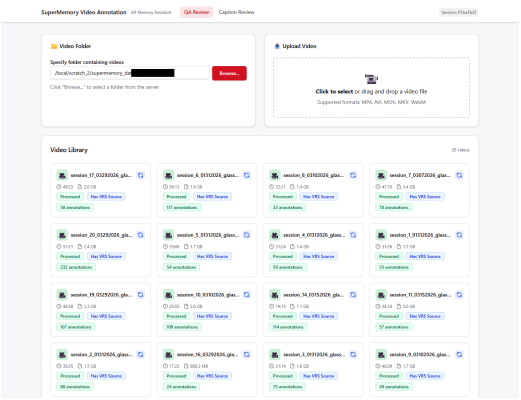
Primary reviewers inspected the generated QA candidates against the source videos and evidence timelines, requiring more than 400 hours of manual review. They rejected 23.5% of QA candidates and retained 4,853 QAs. Each retained QA was reviewed by one primary annotator. The primary annotators were session attendants or project reviewers familiar with the corresponding recordings, which helped resolve ambiguous references and intended actions but may also introduce annotator dependence. To quantify reliability beyond this initial review, another pool of reviewers independently re-evaluated each of the 4,853 QA pairs without consulting the primary annotators, yielding 95.3% agreement on factual correctness, 85.2% for whether an evidence time span required revision and 99.5% for whether the QA text required revision. All identified issues were adjudicated and corrected in the revised annotations.

Reviewers checked factual grounding against the source video and evidence timeline, temporal causality, answerability, question naturalness, and answer-choice balance before examples were included in the final benchmark. Reviews were performed by the session attendant or reviewer familiar with the task context, which helped resolve ambiguous references and intended actions. Because this workflow is not an independent user-utility study, we separately use participant survey responses in §5.3 to assess perceived realism and usefulness.

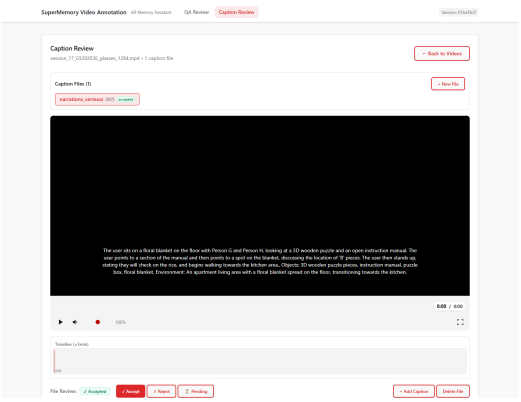
14 Evaluation Protocol and Data Use

The numbers in Table 2 use `SuperMemory-VQA` as an open zero-shot evaluation benchmark. The frameworks and VLMs reported there are not trained or fine-tuned on `SuperMemory-VQA` QA labels; each system receives the question, answer choices, and only evidence available before the question time. We intentionally release labels rather than maintaining a hidden test split, prioritizing reproducibility and flexible research use over leaderboard-style evaluation. The released labels may also support finetuning, retrieval diagnostics, and other supervised analyses, but researchers should clearly state whether a result is zero-shot, finetuned, or otherwise uses `SuperMemory-VQA` supervision.

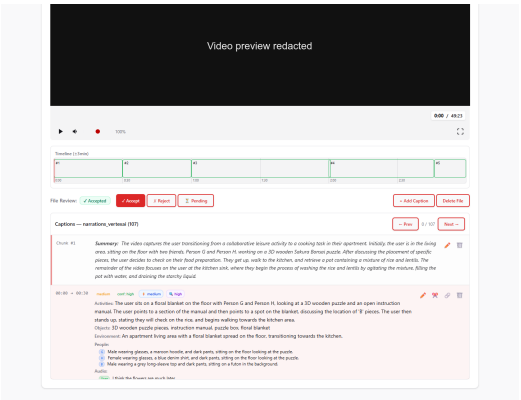
Baseline modality use. The released dataset includes RGB video, transcripts, gaze, motion, trajectory, IMU, and SLAM-derived data, but the current baselines do not exhaust all sensor streams. In practice, Video-RAG mainly consumes RGB frames, ASR transcripts, OCR, object detections, and retrieved auxiliary text, while EgoButler uses RGB/audio-derived clip captions and hierarchical text memories.



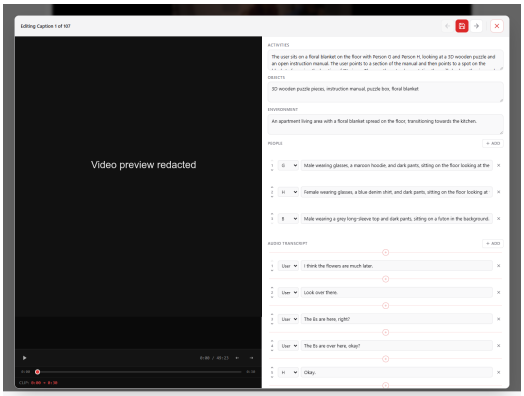
(a) Initial shared review screen.



(b) Caption file and local timeline.

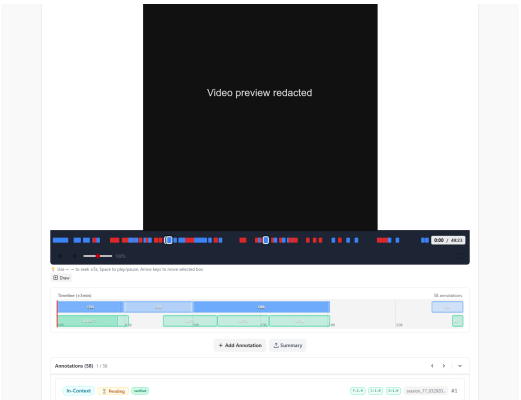


(c) Caption chunks and review controls.

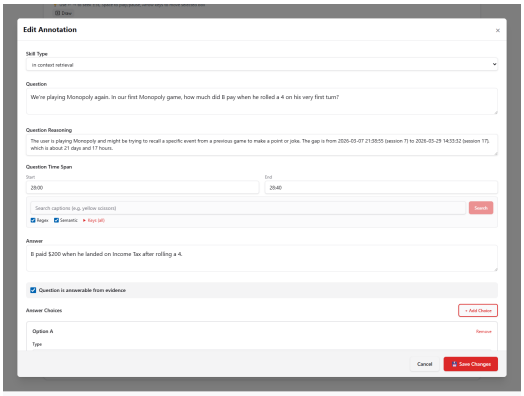


(d) Caption edit screen.

Figure 13: Caption-review flow after Phase 1. Reviewers select a video, inspect caption files and temporal chunks, assign file-level review status, and edit structured caption fields. Video previews are redacted in the figure for privacy.



(a) QA review timeline and annotation panel.



(b) QA edit screen.

Figure 14: QA-review flow after Phase 2. Reviewers inspect temporal evidence markers and generated question-answer annotations, then edit task labels, question text, reasoning, answer choices, evidence, and human-review status before saving. Video previews are redacted in the figure for privacy.

Thus Table 2 should be read as a strong baseline suite for current long-video memory agents, not as an upper bound on methods that explicitly exploit gaze, trajectory, IMU, or SLAM. Currently, there are no agentic system baselines that fully utilize all these modalities.

15 Baseline Implementation

Video-RAG. Using the authors' code, session histories are partitioned into 30-minute shards with per-shard FAISS indices; at query time, the retrieval request fans out across all shards preceding the question, and the top-scoring auxiliary texts are merged. The VLM receives 32 uniformly sampled frames from the most relevant shard plus the merged texts.

EgoButler. EgoGPT is replaced with the base VLM in each row of Table 2. Clip captions are generated over 30-second windows at 1 fps, using WhisperX transcripts for the audio channel. Hour- and day-level summaries are produced by Gemini-3-flash-preview.

WorldMM. We use the official WorldMM implementation (Yeo et al., 2026) with its default retrieval models (Qwen3-Embedding-4B (Zhang et al., 2025) for text, VLM2Vec-V2 (Meng et al., 2025) for video) and at most five retrieval rounds. The 3-min, 10-min, and 1-h episodic summaries and the semantic triples are built by the same backbone that answers the questions.

16 Additional Results

This section complements the main benchmark results in §5.1 and the detailed analysis in §5.2.

Performance degrades as evidence moves farther back in time. Figure 15 analyzes how model performance changes as the temporal distance between the query and its supporting evidence increases. Long-horizon recall remains a central source of difficulty in `SuperMemory-VQA`: systems are more reliable when evidence is temporally nearby, but performance degrades as the relevant memory must be recovered from farther back in the recording history. This trend complements the dataset-statistics discussion in §3.5 and the temporal-gap statistics in Appendix 13, highlighting why egocentric memory benchmarks should evaluate retrieval over realistic time spans rather than only short video clips.

Compositional evidence increases retrieval and reasoning difficulty. Figure 16 evaluates performance as the amount of evidence needed to answer a question increases. Questions grounded in a small number of evidence segments are easier to answer, while questions requiring aggregation across multiple pieces of evidence expose larger failures in retrieval, grounding, and cross-event reasoning. This confirms that `SuperMemory-VQA` stresses not only whether a model can find an isolated relevant moment, but also whether it can compose evidence across a longer memory trace.

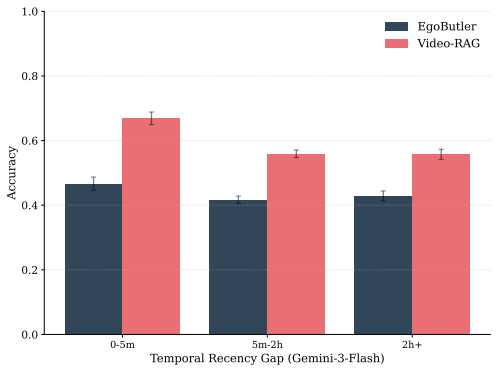


Figure 15: Temporal robustness analysis showing how model performance varies with the time gap between the query and the evidence required to answer it.

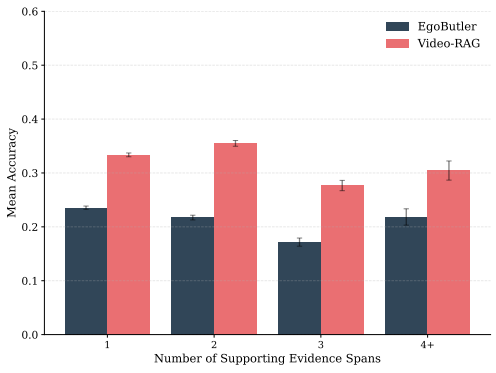


Figure 16: Complexity scaling analysis showing how model performance changes as questions require more supporting evidence.

16.1 Comparison with VideoAgent

We additionally evaluate VideoAgent (Fan et al., 2024) on `SuperMemory-VQA`. Due to its iterative agentic loop, VideoAgent consumes substantially more tokens and computation than EgoButler and Video-RAG, while yielding less competitive results. We therefore report VideoAgent using Gemini-3-Flash only, which offers a favorable cost-performance trade-off for this comparison. As shown in Figure 17, VideoAgent underperforms both EgoButler and Video-RAG across most dimensions, with the largest gaps on Conversational Memory, Intent Recall, and Timeline Reconstruction. The three methods perform comparably on Object Location Memory, indicating that this dimension is less sensitive to the choice of retrieval or agentic strategy. Overall, the added computational cost of VideoAgent does not translate into improved performance on `SuperMemory-VQA`, motivating its exclusion from the main comparison.

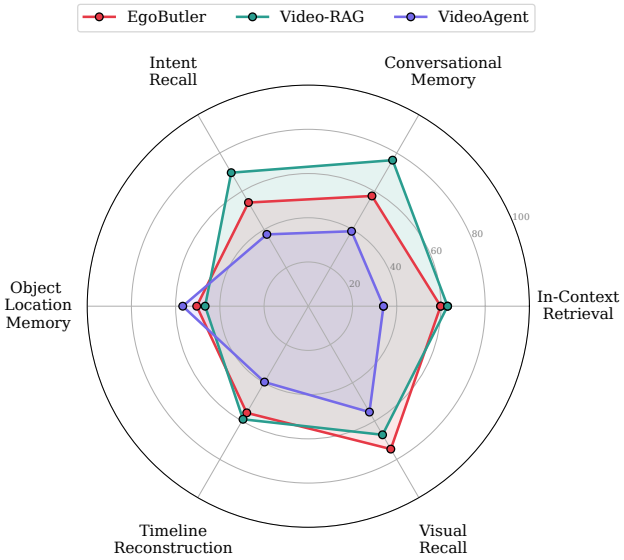


Figure 17: Performance of EgoButler, Video-RAG, and VideoAgent using gemini-3-flash model across six `SuperMemory-VQA` task categories.

16.2 Blind Text-Only LLM Evaluation

A common concern in multiple-choice video QA benchmarks is *information leakage*: questions may be answerable from linguistic priors alone, without genuine grounding in visual, auditory, or temporal evidence. Surface cues in question phrasing, implausible distractors, or world knowledge encoded in large language models can yield non-trivial accuracy and inflate apparent benchmark difficulty. To quantify this risk, we conduct a *blind* evaluation in which a strong LLM receives only the question text and answer options.

We evaluate Gemini-3.5-flash in a text-only configuration. Each prompt contains a brief instruction, the question stem, and four options labeled A–D; no frames, audio, transcripts, or metadata are provided. We report the overall accuracy and accuracy by task category. Table 5 reports results on all subjects ($N = 4{,}853$ questions). Overall accuracy is 31.0%, slightly above the 25% four-way chance baseline, which exhibits minimal information leakage.

16.3 Per-video and per-person Evaluation

Table 6 reports model accuracy on the `SuperMemory-VQA` benchmark across two evaluation granularities. **Per-Video** accuracy is computed per video clip: for each video, we calculate the mean QA accuracy over all questions associated with that clip, then report the mean and standard deviation of these per-video scores across all videos in the benchmark. **Per-Person** accuracy aggregates at the participant level: for each person, we compute the mean QA accuracy over all questions attributed to

Table 5: Blind text-only evaluation with Gemini-3.5-flash. The model receives only question text and options; no visual, audio, or temporal evidence is provided. Chance is 25%.

Category	Correct	Total	Accuracy (%)
Overall			
All questions	1504	4853	31.0
By task category			
conversational_memory	388	824	47.1
object_location_memory	242	824	29.4
visual_recall	146	776	18.8
timeline_reconstruction	243	680	35.7
in_context_retrieval	194	875	22.2
intent_recall	291	874	33.3

that individual, then report the mean and standard deviation across all participants. Both metrics are reported for three retrieval baselines, **Video-RAG**, **EgoButler**, and **WorldMM**, spanning open-source and closed-source vision-language models.

Table 6: Model accuracy on `SuperMemory-VQA` across two evaluation granularities. Mean±StdDev (%) reported; best per column in **bold**.

Model	Per-Video			Per-Person		
	Video-RAG	EgoButler	WorldMM	Video-RAG	EgoButler	WorldMM
Open-source models						
Qwen-3.5-9B	$49.2_{\pm 15.8}$	$39.9_{\pm 12.2}$	$49.8_{\pm 12.4}$	$47.2_{\pm 4.7}$	$40.8_{\pm 6.1}$	$47.9_{\pm 0.8}$
Qwen-3.8-27B	$54.4_{\pm 19.6}$	$41.8_{\pm 15.5}$	$56.7_{\pm 14.5}$	$51.2_{\pm 4.9}$	$41.4_{\pm 5.8}$	$52.9_{\pm 1.7}$
InternVL-3.5 8B	$45.7_{\pm 11.8}$	$39.8_{\pm 12.0}$	$39.9_{\pm 10.9}$	$42.7_{\pm 2.9}$	$41.7_{\pm 5.2}$	$36.6_{\pm 1.7}$
Gemma-4-E4B	$37.6_{\pm 18.8}$	$39.7_{\pm 17.8}$	$50.9_{\pm 12.1}$	$40.2_{\pm 2.4}$	$39.6_{\pm 4.0}$	$49.1_{\pm 1.6}$
Closed-source models						
Gemini-3.8-Flash	$\mathbf{61.5}_{\pm 16.3}$	$\mathbf{64.1}_{\pm 14.3}$	$\mathbf{75.8}_{\pm 11.4}$	$60.4_{\pm 5.2}$	$\mathbf{62.6}_{\pm 5.3}$	$\mathbf{73.0}_{\pm 3.2}$
Gemini-3-Flash	$60.9_{\pm 15.2}$	$59.6_{\pm 13.1}$	$72.8_{\pm 10.5}$	$\mathbf{61.2}_{\pm 6.7}$	$58.3_{\pm 4.8}$	$68.1_{\pm 3.3}$
Gemini-3.1-Pro	$53.0_{\pm 18.4}$	$45.8_{\pm 16.9}$	$72.4_{\pm 7.9}$	$54.5_{\pm 7.4}$	$46.0_{\pm 5.7}$	$70.4_{\pm 3.3}$
GPT-6-sol	$58.2_{\pm 11.2}$	$56.5_{\pm 9.2}$	$71.0_{\pm 7.5}$	$59.9_{\pm 3.5}$	$56.1_{\pm 5.7}$	$70.3_{\pm 2.7}$

Best score in each evaluation-granularity column is highlighted.

Table 7 reports a complementary accuracy variant that assigns partial credit to vague choices. This diagnostic tests whether model differences are driven mainly by choosing vague but related answers instead of fully correct answers. Gemini-3-Flash remains ahead of Gemini-3.1-Pro under both frameworks, indicating that the closed-source scaling gap in the main results is not explained simply by Pro selecting more vague answers.

Table 7: QA average accuracy over all videos when assigning vague choices partial credit of 0.5, along with the absolute change (Δ) from standard QA accuracy where vague choices receive 0 credit.

Model	Video-RAG	EgoButler	WorldMM
Open-source models			
Qwen-3.5-9B	51.3 [↑4.3%]	43.5 [↑2.9%]	52.5 [↑5.9%]
Qwen-3.8-27B	55.7 [↑4.8%]	43.6 [↑2.5%]	57.9 [↑5.5%]
InternVL-3.5 8B	49.1 [↑8.1%]	44.6 [↑4.8%]	43.0 [↑6.6%]
Gemma-4-E4B IT	39.7 [↑2.7%]	38.7 [↑2.2%]	52.6 [↑3.9%]
Closed-source models			
Gemini-3.8-Flash	63.0 [↑2.0%]	**61.0** [↑1.7%]	**72.5** [↑2.2%]
Gemini-3-Flash	**64.4** [↑3.5%]	56.4 [↑2.3%]	68.6 [↑2.3%]
Gemini-3.1-Pro	55.2 [↑2.0%]	43.6 [↑1.1%]	71.0 [↑1.4%]
GPT-6-sol	63.2 [↑3.5%]	58.4 [↑2.2%]	72.6 [↑2.0%]

Best score in each partial-credit QA-accuracy column is highlighted.

Table 8: Retrieval recall and QA accuracy with retrieved or gold evidence. All QA results use Gemini-3-Flash.

Evaluation setting	Recall@5	Recall@30	QA-Acc.
Video-RAG	38.5	70.0	61.0
EgoButler	28.3	60.6	54.1
Gold evidence	–	–	69.7

16.4 Retrieval Recall and Gold-Evidence QA

To distinguish retrieval failures from downstream reasoning failures, we evaluate retrieval Recall@k against the annotated supporting-evidence spans and measure QA accuracy when the gold evidence is provided directly. For each answerable question, Recall@k measures the fraction of annotated evidence spans recovered among the top-k retrieved items, averaged across questions. For the gold-evidence condition, we provide Gemini-3-Flash with the exact ground-truth video segments, together with the question and answer choices, thereby removing the need to search the full recording history.

As shown in Table 8, Video-RAG retrieves a larger proportion of the annotated evidence than EgoButler at both values of k, consistent with its stronger end-to-end QA accuracy. Supplying the gold evidence increases accuracy from 61.0% for Video-RAG and 54.1% for EgoButler to 69.7%. These improvements show that incomplete evidence retrieval is an important source of error. However, accuracy remains below 70% even when the exact supporting clips are provided, indicating that retrieval is not the only bottleneck. Models must still integrate evidence across clips, resolve temporal and entity relationships, distinguish precise answers from plausible distractors, and determine whether the available evidence is sufficient. SuperMemory-VQA therefore exposes failures in both long-horizon evidence retrieval and evidence-grounded reasoning.

16.5 Qualitative Results

To complement the aggregate benchmark results, we inspect a small set of curated examples drawn from the released QA annotations. These examples are intended to explain the concrete reasoning operations that make long-horizon egocentric memory difficult. The main pattern is that failures rarely arise from a single missing capability. Successful answers require the system to retrieve the right moment, preserve the relevant detail through summarization, compare it against the current query, and decide whether the available evidence is sufficient. A model can succeed at one stage and still fail downstream if it loses precision, over-compresses the memory trace, or treats plausible context as evidence.

The inspected examples fall into six recurring qualitative patterns. The figures show the concrete question text, answer choices, model selections, and evidence frames; the discussion below focuses on what each pattern reveals about the benchmark rather than repeating the figure content.

Pinpoint retrieval of sparse evidence. When the decisive evidence is a short recoverable moment, retrieval can answer precisely. Figure 18 shows two cases where the needed information is localized to a few frames or a brief conversational exchange. These examples clarify why retrieval quality matters even when the final reasoning step is simple: once the right evidence is surfaced, the answer is almost direct. The contrast between Video-RAG and EgoButler also suggests that hierarchical summaries can sometimes blur or omit isolated details that are not globally salient, while direct retrieval over lower-level traces can preserve them.

At the same time, these are the easiest qualitative successes: the evidence is sparse but not conceptually complex. They therefore provide a useful positive control for the benchmark. A system that cannot solve these cases is failing at basic memory lookup, whereas a system that can solve them may still struggle on cases that require disambiguation, counting, temporal ordering, or premise checking.

Conversational fact conflation. Conversational memory can fail when nearby context provides a tempting but incorrect paraphrase. In these cases, the model often retrieves the right conversational neighborhood but collapses distinct statements into a broader summary. This is especially problematic for memory assistance because users often ask about small conversational commitments, preferences,

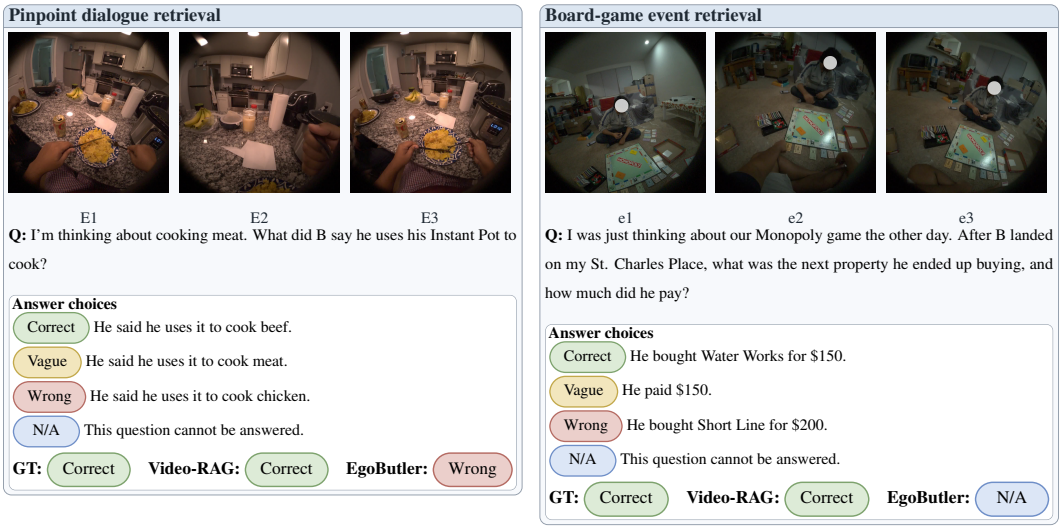


Figure 18: Retrieval cases where brief evidence is sufficient once the relevant moment is found.

or corrections. A vague summary may be semantically related, but it is not the remembered fact the user needs.

This pattern reveals a tension between compression and fidelity. Long-horizon systems need summarization to scale, yet conversational memory frequently depends on preserving exact attributes: who said what, which object was being compared, and whether a statement was a correction or a new fact. The errors here are not random hallucinations; they are plausible substitutions drawn from adjacent context. That makes them hard to catch unless the benchmark distinguishes correct answers from vague or misleading alternatives.

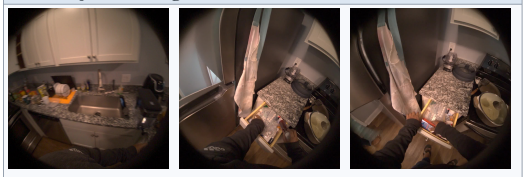
Small visual evidence and OCR-style failures. Fine-grained visual memory is fragile when the answer depends on a small object, a short-lived view, or text that occupies only a small region of the frame. Figure 19 shows two such cases. The important observation is that failure can take two different forms: conservative systems abstain even though evidence exists, while less conservative systems may answer with a visually plausible but incorrect detail. Both are undesirable for an AR memory assistant, but they have different user-facing costs.

These examples also show why egocentric memory cannot be reduced to generic video captioning. A captioner may accurately describe the overall scene while omitting the small object or tiny text that later becomes decisive. Once that information is absent from the memory representation, downstream retrieval has little chance to recover it. Robust systems likely need adaptive high-resolution inspection, object permanence over brief interactions, and OCR that can be triggered by later user intent rather than only by frame-level salience at indexing time.

Temporal ordering and state-change tracking. Long-horizon memory requires preserving order or object state and activity, not merely retrieving a visually similar frame. Figure 20 contrasts a sequencing failure with a state-change success. The hard part is that the relevant frames can look repetitive: similar objects, similar hand motions, and similar locations recur across the session. A memory system must therefore maintain an event-level trace that records transitions, not just a bag of matching visual observations.

The state-change case illustrates what success looks like: the system must link an object's earlier use to a later cleaning event and then to its final storage location. This is closer to episodic state tracking than ordinary visual recognition. The sequencing failure, by contrast, shows that even when all evidence is present, the model may not preserve the ordering relation with enough confidence to answer. These failures point toward architectures that explicitly represent temporal predicates such as before, after, moved-to, cleaned, stored, and still-missing.

Small-object tracking failure



e1 e2 e3

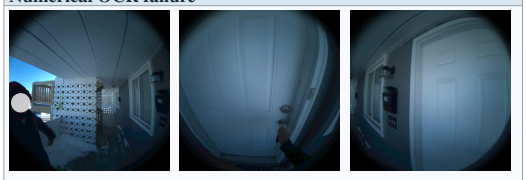
Q: I need to take my medicine before I start cooking. Where did I leave that small white bottle?

Answer choices

- Correct: You left it in the kitchen drawer next to the stove.
- Vague: You left it in one of the drawers in the kitchen area.
- Wrong: You left it in the lower cabinet under the kitchen island.
- N/A: This question cannot be answered.

GT: Correct **Video-RAG:** N/A **EgoButler:** N/A

Numerical OCR failure



e1 e2 e3

Q: I'm trying to order delivery on my phone. What was the apartment number on the door we came through earlier?

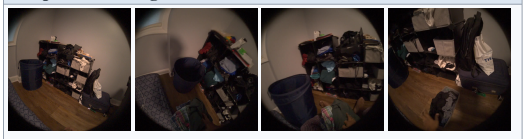
Answer choices

- Correct: The apartment number printed next to the door was 205.
- Vague: The apartment number was 205 or 207.
- Wrong: The apartment number printed next to the door was 207.
- N/A: This question cannot be answered.

GT: Correct **Video-RAG:** Wrong **EgoButler:** N/A

Figure 19: Small visual evidence cases where the decisive cue is either a brief object observation or fine-grained text.

Temporal ordering failure



socks shorts shirt later

Q: I'm trying to remember how I organized my laundry. In what order did I place the black socks, grey t-shirt, and blue patterned shorts into the storage organizer?

Answer choices

- N/A: This question cannot be answered.
- Vague: You put the socks away before the shirts and shorts.
- Correct: You placed the black socks in first, followed by the blue patterned shorts, and finally the grey t-shirt.
- Wrong: You placed the grey t-shirt in first, followed by the black socks, and finally the blue patterned shorts.

GT: Correct **Video-RAG:** N/A **EgoButler:** N/A

State-change tracking success



used washed stored later

Q: I'm loading up the dishwasher right now. Did I already wash the frying pan I used to roast the spices, or do I still need to track it down and put it in here?

Answer choices

- Correct: You already hand-washed the frying pan in the sink and put it away in a lower kitchen cabinet.
- Vague: You already cleaned the frying pan earlier today and put it away in one of the cabinets.
- Wrong: You loaded the dirty frying pan into the bottom rack of the dishwasher, so it is already in there.
- N/A: This question cannot be answered.

GT: Correct **Video-RAG:** Correct **EgoButler:** N/A

Figure 20: Temporal reasoning examples where the answer depends on preserving action order or object state across evidence frames.

Quantifier precision. Repeated near-identical actions expose counting failures. The issue is not only detecting an object category, but deciding whether multiple observations correspond to the same item, repeated handling of one item, or distinct instances. This is a common daily-memory need: users ask whether they already added an ingredient, how many pieces were packed away, or whether a repeated action happened once or several times.

The qualitative count failures suggest that current memory pipelines do not reliably maintain instance identity across repeated actions. Summaries tend to compress repeated visual events into phrases such as "some items" or "several actions," which are useful for gist but insufficient for exact answers. A stronger memory assistant would need explicit count-preserving representations, uncertainty over duplicate observations, and mechanisms for reconciling repeated views of the same object.

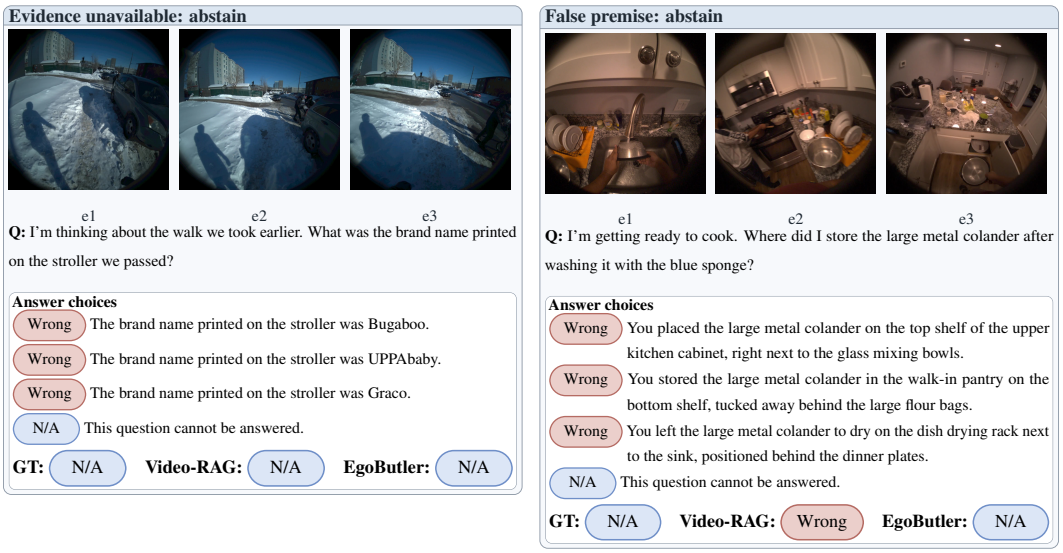


Figure 21: Abstention and premise-validation examples where related evidence exists but the requested fact is unsupported.

Abstention and premise validation. The unanswerable option is essential when the event is partially visible but the requested detail is missing, or when the question contains a false premise. Figure 21 shows both cases. These examples are important because they separate retrieval from epistemic validation: finding a related visual moment is not enough. The model must also verify that the specific requested fact is actually supported.

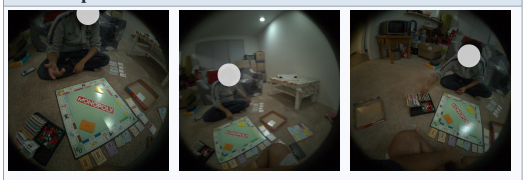
False-premise questions are especially revealing. A system may retrieve the named object and then answer as though every clause in the query were true. From the user's perspective, this is a hallucination even if the object itself was seen. Correct abstention therefore requires checking the full proposition expressed by the question, including modifiers, causal claims, and implied actions. This is why `SuperMemory-VQA` includes an explicit unanswerable option rather than treating all memory queries as answerable lookup problems.

Taken together, these qualitative cases show that long-horizon AR memory requires a coupled solution: retrieval must be sensitive enough to surface sparse details, memory representations must preserve exact attributes and event order, and the answerer must determine what is and is not supported by the evidence. Aggregate accuracy alone hides these distinctions. The qualitative breakdown makes clear that future progress will likely require explicit memory structures for object state, counts, temporal relations, and premise validation, in addition to stronger vision-language models.

Gemini-3-Flash versus Gemini-3.1-Pro. The results expose a sharper model-level trade-off between Gemini-3-Flash and Gemini-3.1-Pro under the same retrieval setting. Figure 22 shows answerable cases where Flash is willing to use retrieved evidence, while Pro abstains despite support being present. In the Monopoly arithmetic case, the model must retrieve two distinct rent events and compute the difference. In the card-text case, the model must read a small piece of game text and

preserve the exact wording. These examples are consistent with the aggregate result that Flash often obtains higher QA accuracy by committing when evidence is available.

Multi-step arithmetic retrieval



rent later calc

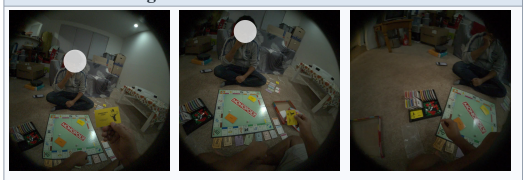
Q: Hey Aria, thinking back to our Monopoly game, what was the difference between that $200 rent I paid B and the rent he paid me for St. Charles Place?

Answer choices

- Correct: You paid B $186 more than he paid you for St. Charles Place.
- Vague: You paid B significantly more than he paid you for St. Charles Place.
- Wrong: You paid B $150 more than he paid you for St. Charles Place.
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Small-text reading



card text later

Q: I'm texting my friend to settle a debate about our Monopoly game from the other day. What was the exact text on that Community Chest card I drew that gave me 50 dollars?

Answer choices

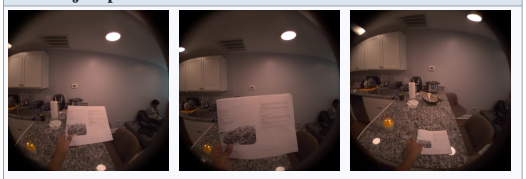
- Correct: The exact text on the card was "FROM SALE OF STOCK YOU GET $50."
- Vague: The text on the card mentioned receiving $50 from a stock sale.
- Wrong: The exact text on the card was "BANK ERROR IN YOUR FAVOR GET $50."
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Figure 22: Answerable Flash and Pro comparisons under Video-RAG. Flash commits to retrieved evidence, while Pro abstains on multi-step arithmetic and small-text recall.

The same tendency reverses on false-premise questions. Figure 23 shows cases where the retrieved context is topically related but does not support the proposition in the question. Here, Flash produces specific but unsupported answers, whereas Pro correctly abstains. This suggests that Pro's conservatism can be useful when the evidence is missing or the prompt smuggles in an unverified event, even though the same conservatism hurts answerable recall.

False object premise



recipe counter later

Q: I'm looking for a tool to help with this recipe. Where did I leave the blue measuring cup?

Answer choices

- Wrong: The blue measuring cup is inside the dishwasher, which is currently closed.
- Wrong: The blue measuring cup is in the microwave cabinet, tucked behind some plates.
- Wrong: The blue measuring cup is on the counter, hidden behind the dish drying rack.
- N/A: This question cannot be answered.

GT: N/A **Gemini-3-Flash:** Wrong **Gemini-3.1-Pro:** N/A

False temporal premise



game floor later

Q: Now that we're sitting down to play Monopoly, did I actually pick up the Monopoly Deal box before I suggested we play it, or was it the other way around?

Answer choices

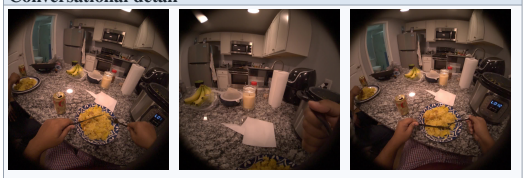
- Wrong: You picked up the Monopoly Deal card game box first, and then suggested playing it later.
- Wrong: You suggested playing Monopoly Deal first, and then picked up the game box to show it.
- Wrong: You picked up the box and suggested playing it while clearing the living room floor.
- N/A: This question cannot be answered.

GT: N/A **Gemini-3-Flash:** Wrong **Gemini-3.1-Pro:** N/A

Figure 23: Unanswerable Flash and Pro comparisons under Video-RAG. Pro correctly rejects unsupported object and temporal premises, while Flash gives plausible but fabricated details.

Together, these additional examples explain why the Flash/Pro comparison is not a simple model-scale story. Flash is often better matched to answerable retrieved evidence, especially when the task requires committing to small visual text, conversational facts, object placement, arithmetic over events, or procedural order. Pro is more cautious and can therefore lose recall, but that caution protects against false positives when the evidence does not actually support the user's premise.

Conversational detail



talk meal later

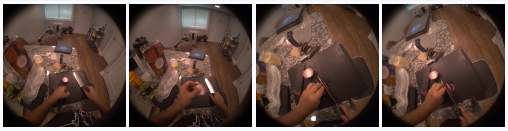
Q: What did B say he uses his Instant Pot for when we were eating biryani?

Answer choices

- Correct: He said he uses it for cooking beef, because it makes the meat really soft compared to the stovetop.
- Vague: He said he uses it for cooking meat, because it produces a better texture than when using the stovetop.
- Wrong: He said he uses it for cooking chicken, because it cooks the meat much faster than on the stovetop.
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Procedural sequence



ends peel half slice

Q: I'm writing down the recipe steps for a friend. How exactly did I prep the red onion earlier?

Answer choices

- Correct: Ends off, outer layers peeled, onion cut in half, then sliced into thin strips.
- Vague: Ends and skin removed, then the onion was cut smaller and sliced into strips.
- Wrong: Outer layers peeled first, then the ends were sliced off before cutting and slicing.
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Figure 24: Additional answerable Flash and Pro comparisons under EgoButler. Flash uses summarized conversation and procedural logs, while Pro abstains.

Package-label recall



salt label later

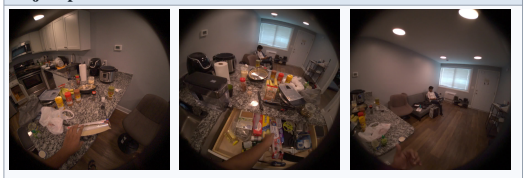
Q: I'm checking the spices. Does the blue Morton salt container say if the salt is iodized?

Answer choices

- Correct: Yes, the blue Morton salt container indicates that the salt is iodized.
- Vague: The label on the blue container provides details about the salt type.
- Wrong: No, the blue Morton salt container indicates that it is plain sea salt.
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Object placement



box drawer closed

Q: I need to wrap up some leftovers. Where did I put the box of plastic wrap?

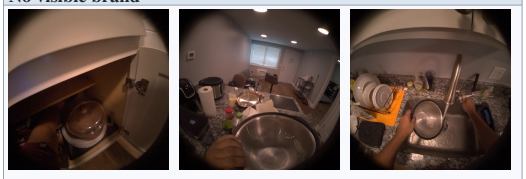
Answer choices

- Correct: You stored the box of plastic wrap in a lower drawer in the kitchen island.
- Vague: You put it away in one of the kitchen drawers.
- Wrong: You stored the box of plastic wrap in the upper cabinet above the microwave.
- N/A: This question cannot be answered.

GT: Correct **Gemini-3-Flash:** Correct **Gemini-3.1-Pro:** N/A

Figure 25: Additional answerable Flash and Pro comparisons under Video-RAG. Flash answers small-label and object-location questions, while Pro abstains.

No visible brand



strainer colander sink

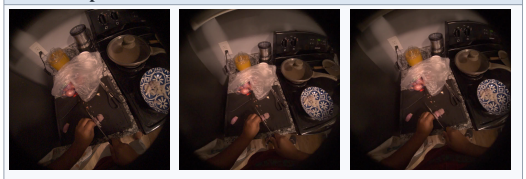
Q: Hey, I remember using a white mesh strainer and a metal colander earlier today. Was either of those made by Cuisinart?

Answer choices

- Wrong — The white mesh strainer is a Cuisinart model, but the brand of the metal colander is not visible.
- Wrong — The metal colander is manufactured by Cuisinart, while the brand of the white mesh strainer is not recorded.
- Wrong — Both the white mesh strainer and the metal colander are identified as Cuisinart products.
- N/A — This question cannot be answered.

GT: N/A **Gemini-3-Flash:** Wrong **Gemini-3.1-Pro:** N/A

False tool premise



lentils spoon prep

Q: I'm cleaning up my prep station. Where did I put the wooden spoon I used to scoop the lentils?

Answer choices

- Wrong — You used a silver spoon to scoop the lentils, not a wooden spoon.
- Wrong — You used a white spoon to scoop the lentils, not a wooden spoon.
- Wrong — You used a blue spoon to scoop the lentils, not a wooden spoon.
- N/A — This question cannot be answered.

GT: N/A **Gemini-3-Flash:** Wrong **Gemini-3.1-Pro:** N/A

Figure 26: Additional unanswerable Flash and Pro comparisons under EgoButler. Pro rejects noisy summary evidence and false tool premises more reliably than Flash.