# NEVER LOOK BACK: UNDERSTANDING PERSISTENCE IN 3D OBJECT MEMORY FROM EGOCENTRIC VIDEOS

**Anonymous authors**
Paper under double-blind review

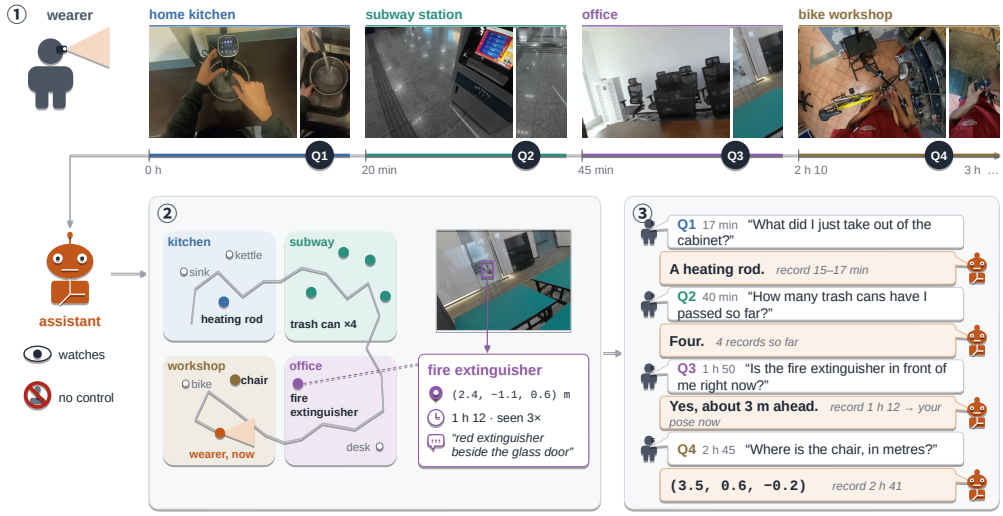


Figure 1: **Persistent object memory from egocentric video.** The illustration shows ① A wearer/assistant observes a continuous stream and performs several tasks for hours in a scene (or spanning multiple scenes), the questions about which arrive at different times. ② An assistant (physical or digital) keeps a record for each object it (or the wearer) sees, with a 3D position, the times it was seen, and a short description, placed on a map of the path traveled. ③ Later questions are answered from those records alone, without replaying the video: recalling an object (Q1), counting (Q2), a direction relative to where the wearer stands now (Q3), and a metric position (Q4).

## ABSTRACT

As we move through the world and carry out everyday tasks, we encounter objects that may become relevant only later. We are capable of recalling where we left something or what was inside a container, even without knowing we would need it later. Here, we study how an embodied assistant can build a similar memory from egocentric videos, by observing a person's day-to-day activities. We present LEDGER, a persistent 3D object memory that combines object locations, their histories, and contextual descriptions. It associates observations across the recording and retains objects after they leave the view, including those the person never touches. It clusters each object's observations by resting locations and records a move only after repeated evidence, reducing the effect of localization noise. Short descriptions preserve details such as an object's contents or supported surface. It saves these records to later answer spatial questions without having to access the original images or video. Our memory raises HD-EPIC accuracy from 29.7% to 42.6%, UCS-Bench accuracy from 33.8% to 38.5% and localizes Ego4D objects with a 0.99 m median error on returned predictions. Our analyses identify complementary roles for temporal persistence, contextual descriptions, and retrieval. Our study on 100 stitched streams of multiple scenes each further exposes failures in both retrieval and construction. Per-scene construction partially recovers the performance lost across scene changes compared to that of single scene streams.

Code is available here, and visualizations are included in supplementary materials with a README file.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Humans typically perform several tasks during the day like cooking, clearing out the garage, playing football, or other diverse activities. We are reliably good at tracking our environments and our positions relative to the objects of the scene we are in (Wang & Spelke, 2000; Burgess, 2006). Even after exiting the scene, we can still answer certain questions about it like where we left the wrench, which shelf had the spare batteries, or where was the serving bowl last placed (Postma et al., 2008; Hollingworth & Henderson, 2002). Fundamentally, we keep a running memory of where potentially important objects are (Land et al., 1999; Helbing et al., 2020), freeing us from having to exactly retrace our steps in either our memories or the scene itself (Tolman, 1948; Epstein et al., 2017). Similarly, in long horizon explorations spanning multiple scenes, we are capable of both tracking objects across scenes (Warren et al., 2017) and revising our memory for familiar objects displaced when we were absent from a scene (Hollingworth, 2005; Hupbach et al., 2008). Our goal is to develop embodied spatial memory that supports recall of previously visited scenes and incorporates new observations of objects and changes in familiar environments. A natural setting to study these capabilities are long egocentric videos about everyday human activity. Such videos capture people moving through diverse scenes, interacting with objects, and revisiting places over time. This enables us to study what an embodied agent can retain from a person's visual experience without controlling where that person looks or what they do.

Building such a memory from these recordings is challenging. Objects can be occluded, briefly visible or captured from an alternate viewpoint (Grauman et al., 2022; Damen et al., 2022; Tang et al., 2023). They may also move between observations, and errors in 3D localization can make an agent mistake a stationary object as moving. Moreover, the agent does not know during exploration which objects or events will be relevant to a question asked later. It should decide what information to retain from these incomplete observations. This includes information about objects that are only seen in passing. In this work, we study four aspects of this problem: which objects are recorded in the memory, how to record changes in their locations, what information is retained beyond their 3D positions, and how earlier observations can be retrieved as more scenes are added. Our goal here is to understand how these design choices impact the spatial questions an agent can answer after exploration, without revisiting the scene or replaying the video.

Recent works have developed object memories that retain identities and locations after objects leave the view (Plizzari et al., 2024; Chalk et al., 2026; Goletto et al., 2024; Manigrasso et al., 2024). Object-centric maps and scene graphs organize observations through the spatial relationships between objects (Gu et al., 2024; Schmid et al., 2024). Language-model agents write their observations into a structured record of objects, places, and their states, which the model later reads to answer questions (Anwar et al., 2024; Fan et al., 2025; Yu et al., 2026; Wang et al., 2026a). Table 1 compares these memories. Another line of work simply uses video-language models to answer from frames sampled from the recording at question time without explicitly maintaining a record of the scene (Cheng et al., 2024b; Zhang et al., 2025b;c; Team et al., 2024), including models trained particularly for spatial reasoning (Wu et al., 2025; Yang et al., 2025b). Streaming models compress past frames in a memory of visual tokens (Zhang et al., 2025a; Qian et al., 2025; Huang et al., 2025). Visual query localization methods search the recording again for every query (Grauman et al., 2022; Mai et al., 2023). To guide memory design, we need to decide which retained evidence supports later questions and to distinguish failures of representation from retrieval and reasoning failures.

For example, restricting memory to objects involved in hand interactions (Goletto et al., 2024; Plizzari et al., 2024) can exclude visible objects that still affect the agent or are relevant later on. Although retaining an object's position may support a location question, it does not preserve its contents or its movement history. Even a sequence of positions can mislead when localization noise is recorded as movement. Further failures can arise across scenes when storing observations into memory or while retrieving relevant records. These examples motivate us to address our core question: what must an agent preserve from past observations to answer spatial questions it could not anticipate?

To this end, we present LEDGER (Long-horizon Egocentric Descriptions, Geometry, and Event Records), a persistent 3D object memory that combines geometric observations with contextual descriptions. We construct the memory offline from sampled egocentric recordings, using supplied or estimated camera poses. We detect objects with an open-vocabulary detector, associate their ob-

Table 1: What existing memories built from egocentric or embodied video keep, and what they need at query time. ✓: supported and demonstrated; ○: partly or only under conditions; ✗: not supported; N.A.: not applicable. The memories differ in their objects' coverage, tracking of motion history, input conditions for construction or handling of scene changes, and query time support for no frame access, language queries and 3D location requests.

| | Memory content | | | | Construction conditions | | | Query time | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | Handled objects | Untouched objects | Motion history | Path & places | RGB only | Scene changes | ≥20-min video | No frames at query | Language queries | 3D location |
| ConceptGraphs (Gu et al., 2024) | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | N.A. | ✓ | ✓ | ✓ |
| 3D-Mem (Yang et al., 2025c) | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | N.A. | ✗ | ✓ | ✓ |
| ReMEmbR (Anwar et al., 2024) | ✗ | ✗ | ✗ | ✓ | ✗ | N.A. | ✓ | ✓ | ✓ | ○ |
| VideoAgent (Fan et al., 2024) | ○ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| AMEGO (Goletto et al., 2024) | ✓ | ✗ | ○ | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| OSNOM (Plizzari et al., 2024) | ✓ | ✗ | ✓ | ✗ | ○ | ✗ | ✓ | N.A. | ✗ | ✓ |
| Emb. VideoAgent (Fan et al., 2025) | ✓ | ✓ | ○ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| EgoRAG (Yang et al., 2025a) | ○ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| VL-KnG (Al Mdfaa et al., 2025) | ✗ | ✓ | ○ | ✓ | ✓ | ✗ | ✓ | ○ | ✓ | ✗ |
| DirectMe (Wang et al., 2026a) | ○ | ✓ | ○ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| R4DSG[e] (Ma et al., 2026) | ○ | ○ | ✓ | ○ | ✓ | ○ | ✓ | ✓ | ✓ | ✗ |
| **LEDGER (ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ○ | ✓ | ✓ | ✓ | ✓ |

[e] R4DSG's objects enter through a per-segment controlled vocabulary and its places through location segments, both taken from released segment records whose source the paper does not state; anchors are found per location window, and the authors report limited cross-day persistence.

servations over time, store them whether or not the agent interacts with them, and keep their records after they leave view. We group each object's timestamped 3D observations into *rest segments*. Each rest segment for an object represents its observations close to the same resting location. A new segment indicates that the object has moved, which is only recorded when multiple subsequent observations support it. We use a persistence rule to reduce false motion detections caused by localization noise. We add a short contextual description to each segment to capture details such as what a container held or which surface it rested on. We first identify the objects referred to in the question and retrieve their records. A language model then answers using these records as text, without seeing images or the original video. This tests whether the stored locations, histories, and descriptions provide enough information to answer questions about objects no longer in view.

We evaluate LEDGER on all 2,400 questions in HD-EPIC's 3D Perception and Object Motion categories (Perrett et al., 2025), named-object localization on Ego4D VQ3D (Grauman et al., 2022), and 2,771 timestamped questions on UCS-Bench (Wang et al., 2026a). On HD-EPIC, adding our memory raises GPT-5.4's[1] prototype-averaged accuracy from 29.7% to 42.6%. On UCS-Bench, reading the memory alone raises accuracy from 33.8% to 38.5%, against 30.0% for DirectMe's graph. On VQ3D, our memory returns object locations with a median error of 0.99 m among the queries for which it returns a prediction.

**Contributions.**

- **A reusable 3D memory of past observations.** We introduce LEDGER, a persistent object memory built offline from egocentric videos with supplied or estimated poses, retaining locations, movement histories, and contextual descriptions, even for untouched objects.
- **Spatial answering and localization across benchmarks.** We demonstrate improved text-only question answering on HD-EPIC, named-object localization on Ego4D VQ3D, and gains over frame-only answering on UCS-Bench, and compare against rebuilt object, caption, and graph memories to show where our representation helps and what it misses.
- **Evidence for how to build and read memory.** Ablations isolate temporal persistence, contextual descriptions, and geometry, and shared readers reveal how retrieval shapes a representation's apparent advantage; costlier iterative search does not improve every task.
- **An analysis of accumulating experience.** Across recording lengths and 100 paired stitched streams, memory retains its gains up to 75 minutes, while scene changes expose failures in construction and retrieval that building memory per scene partially repairs.

[1] GPT-5.4 runs without extended reasoning (thinking) in all our experiments.

## 2 Related Work

**Memories of objects and scenes.** OSNOM tracks the 3D locations of active objects through absence with appearance and spatial association (Plizzari et al., 2024), Whareformer learns association with an updatable track memory and an explicit new-track decision and is trained on active objects. However, its architecture does not require contact (Chalk et al., 2026). AMEGO builds location memories and interaction tracklets for querying on long videos (Goletto et al., 2024), and ESOM couples discovery, tracking, and retrieval in an online visual-query memory (Manigrasso et al., 2024). These systems determine which objects enter memory and how to recognize them when they reappear. Scene representations also differ in the information they store. ConceptGraphs builds open-vocabulary scene graphs of objects from posed RGB-D observations (Gu et al., 2024). Snapshot and scene-based memories support spatial reasoning and exploration (Yang et al., 2025c; Lu et al., 2026; Zhou et al., 2026). Khronos records short-term motion and long-term scene changes (Schmid et al., 2024). DirectMe tracks objects, places, and the wearer, and uses geometry to derive spatial relations for timestamped questions (Wang et al., 2026a).

We study how the objects retained, their estimated positions, movement histories, and contextual descriptions help answer different spatial questions. Errors in geometry can misplace an object or make it appear to have moved. We therefore evaluate how these errors affect answers, alongside their effect on localization.

**Reading stored memories.** Memory-based systems differ in what they store, how they find relevant information, and whether they show images to the answerer. ReMEmbR stores captions with timestamps and positions, then searches this memory over multiple rounds to answer a question (Anwar et al., 2024). VideoAgent and Embodied VideoAgent combine stored information about entities with retrieved frames (Fan et al., 2024; 2025). MEMORA maintains entity state for later reasoning (Yu et al., 2026), while DirectMe gives the answerer its graph and retrieved frames (Wang et al., 2026a). Video-language and streaming models use sampled frames or compressed visual tokens without explicit object records (Cheng et al., 2024b; Zhang et al., 2025b;c; Team et al., 2024; Zhang et al., 2025a; Qian et al., 2025; Huang et al., 2025). We test what stored records can support when the answerer receives only text. We also compare our object records with event descriptions against a caption memory using the same retrieval procedures. This helps distinguish the value of the stored information from the reader's ability to find and use it.

**Evaluating spatial memory over time.** Existing benchmarks test different demands on memory. Ego4D VQ3D queries an object's location using an image crop (Grauman et al., 2022); EgoLoc, ReLoCATE, and EAGLE address this task through pose recovery, training-free retrieval, and appearance and geometric memory (Mai et al., 2023; Khosla et al., 2025; Cao et al., 2025). We evaluate localization using object names instead. HD-EPIC tests object and fixture positions and motion on calibrated recordings (Perrett et al., 2025), while UCS-Bench tests timestamped spatial reasoning relative to the moving wearer (Wang et al., 2026a). EASG-Bench tests questions grounded in scene graphs (Rodin et al., 2025), while OpenEQA tests questions about previously observed environments (Majumdar et al., 2024). HourVideo evaluates reasoning over long egocentric videos (Chandrasegaran et al., 2024). EgoMemReason extends this to evidence spanning days, including changes in object states (Wang et al., 2026b). We complement these benchmarks by asking the same questions on individual recordings and on streams formed by joining recordings. We then modify memory construction and retrieval to identify why performance changes.

## 3 Method

LEDGER constructs a persistent object memory offline from an egocentric recording and supplied or estimated camera poses (Figure 2), once per recording and independently of any question. It is then reused for spatial question answering and named-object localization; on UCS-Bench, reading it costs about 2.0k input tokens per question, against 9.6k for 48 sampled frames (Appendix H).

For each instance $j$, we store its label $c_j$ and timestamped observations $\mathcal{O}_j = \{(t_i, b_i, \mathbf{x}_i)\}$, where $b_i$ is the image box and $\mathbf{x}_i$ is the estimated 3D position in scene coordinates. We summarize the object's observed location history as *rest segments*, $\mathcal{S}_j = \{(t_s^{\text{start}}, t_s^{\text{end}}, \boldsymbol{\mu}_s, d_s)\}$. Each segment groups observations associated with one resting location: $t_s^{\text{start}}$ and $t_s^{\text{end}}$ delimit the segment's span

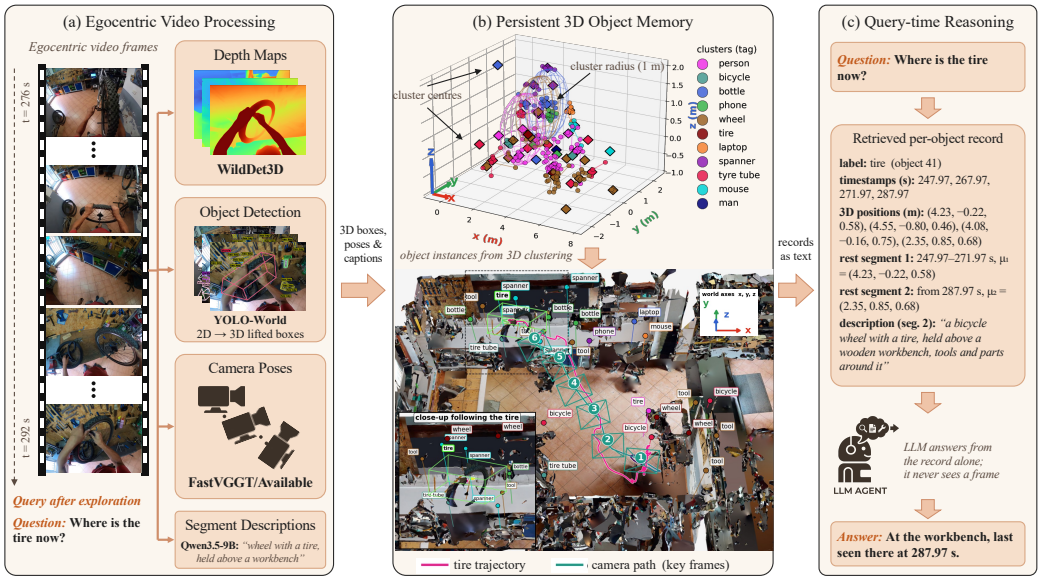


Figure 2: LEDGER overview. (a) Objects detected in sampled egocentric frames are localized in 3D and placed in a scene coordinate frame using supplied or estimated camera poses. (b) Persistent object records, formed from associated observations, contain timestamped positions, rest segments defined by repeated evidence of displacement, and contextual descriptions. (c) After query grounding and retrieval, a language model answers from the records rendered as text, without receiving images. When pose scale is available, scene coordinates are metric.

in the sampled observation sequence, $\boldsymbol{\mu}_s$ represents its estimated 3D location, and $d_s$ describes the object and its surrounding context. A new segment is confirmed only after repeated observations support a displacement from the current location. Rest segments thus summarize where the object was observed; its position during gaps between observations is unknown. Camera poses also retain the wearer's trajectory and heading for egocentric queries. Unrelated scenes are not registered into a shared frame; coordinates are metric where pose scale is available.

### 3.1 OFFLINE MEMORY CONSTRUCTION

**Discovery and geometry.** Qwen3.5-9B names objects in sampled frames; synonym merging produces a vocabulary for YOLO-World (Cheng et al., 2024a), so objects enter memory whether or not the wearer touches them, within the detection and 3D-lifting budgets. We detect with SAM 3 (Carion et al., 2025) on VQ3D, where localization accuracy is what we measure, and with YOLO-World elsewhere for its lower cost. From the image, detection box, and label, WildDet3D (Huang et al., 2026) predicts a 3D box in camera coordinates, and the camera pose places its centre in scene coordinates (Appendix A).

**Association and persistence.** We match observations in time order to existing instances by label and distance from the instance's latest position; same-frame detections stay separate unless they are overlapping duplicates. A second pass re-identifies objects that 3D association split into several tracks: the SAM 3 video tracker (Carion et al., 2025) propagates each fragment's masks and merges fragments whose masks agree, subject to co-occurrence and 3D-distance checks (Appendix A.3). Records persist after objects leave view, although a large move may create a new identity. A new rest segment starts only after $k$ successive observations lie more than $\tau$ from the current resting position; observations that return within this radius stay in the current segment. This suppresses false moves from localization noise but may miss brief ones. **Context and refinement.** Qwen3.5-9B describes selected segments from frames with the object outlined, recording contents, supporting surfaces, and interactions (on HD-EPIC, transferred from distance-threshold segments; Appendix A.4). Multi-view triangulation refines a resting object's position when the viewpoints are well separated and the

estimate is stable and close to the original (Appendix A.5). Two extensions add activity descriptions for sampled frames and build a separate memory for each detected scene (Appendices E.5 and E.4).

### 3.2 QUERY-TIME REASONING

**Grounding and retrieval.** HD-EPIC box queries are lifted to a 3D point on the sampled frame nearest their timestamp, and the four nearest instances observed within $\pm 6$ s are retrieved; named fixtures are matched through fixed synonym groups, taking the most-observed matches, with clock directions computed from stored positions and the camera heading at query time. VQ3D supplies an object name, and we return the selected instance's stored position; UCS-Bench targets are inferred from the question, and entries are ranked by word overlap and temporal proximity (Appendices B.2, C, and D). **Answering.** Retrieved labels, ranking signals, and segment histories (times, positions, and descriptions) are rendered as text, and a single call to Qwen3.5-9B (Team, 2026) or GPT-5.4 (OpenAI, 2026a) answers from this evidence, the question, and its options. HD-EPIC queries use the completed memory; for timestamped UCS-Bench queries, snapshots remove later observations and descriptions, truncate histories and the wearer path, and recompute positions. This filters what the answerer sees without making construction causal (Appendix D.3).

## 4 EXPERIMENTS

Five research questions connect benchmark results to interventions on the memory; Appendix G.1 analyses failures by stage.

**Benchmarks and baselines.** We use all 2,400 3D Perception and Object Motion questions of HD-EPIC (Perrett et al., 2025) and use prototype-averaged accuracy (Appendix B.1). We also use 164 Ego4D VQ3D validation queries (Grauman et al., 2022), localized using the object's name and scored over predicted answers alongside coverage. Finally, we use 2,771 UCS-Bench questions (Wang et al., 2026a) from 211 EgoLife and TeleEgo recordings, with estimated poses and memories filtered to the question time (Appendix D.1). As baselines, we run blind control (question and options only), memory alone, 48 frames sampled up to the question time, and frames plus memory on spatial questions. DirectMe, ReMEmbR, OSNOM, and AMEGO are rebuilt from released code with dataset adaptations and read/answered by GPT-5.4 unless stated; RQ4 examines their retrieval. We further include evaluations on streams spanning multiple scenes Figure 3. It follows DirectMe, ReMEmbR, and LEDGER through one stream. LEDGER is correct on all five questions, DirectMe's graph on one, and ReMEmbR on none, and the difference follows from what each memory keeps. For Q2, the detector's label is wrong, but the description stored with the object's rest segment names what the wearer was holding, which the answerer follows; neither baseline stored the object.

**RQ1: What can object memory answer, and what evidence does it miss?** *The memory helps every answering model we tested, especially on objects the wearer never touched and least on interactions and counts.* On HD-EPIC, the memory lifts Qwen3.5-9B, GPT-5.4, and GPT-5.6 (OpenAI, 2026b) by 10.8 to 17.8 points over their blind controls and, read by GPT-5.4, outscores every rebuilt memory (Table 2; Figure 4b); on UCS-Bench, it also improves the 48-frame reader it is added to (Table 3). The gain is evident for fixture location, because admitting untouched objects supplies evidence that contact-gated memories exclude. The weakest cases need events and identities rather than more objects: AMEGO's hand–object tracklets remain better on interaction counting and the onset of rest (Table 10), and UCS-Bench object counts do not improve because association both splits single objects into several tracks and misses others (Appendix D.2).

**RQ2: When does localization noise get confused for movement?** *Requiring repeated evidence of a move, instead of raising the distance threshold, keeps localization noise from being recorded as movement.* Since they overlap in size, a threshold alone cannot separate the two. Half of the annotated moves are shorter than 0.8 m, and the 95th percentile of localization error is 0.84 m. A 0.3 m threshold (roughly doubles the counted moves), and 0.8 m gets only 57.3% of counts within $\pm 1$ of the truth. Keeping 0.3 m but requiring three consecutive displaced observations raises this to 67.4% (Table 16), and removing persistence costs movement counting 15.5 points, which survives correction for multiple tests (Figure 4a). The price we pay here is that brief moves go unrecorded.

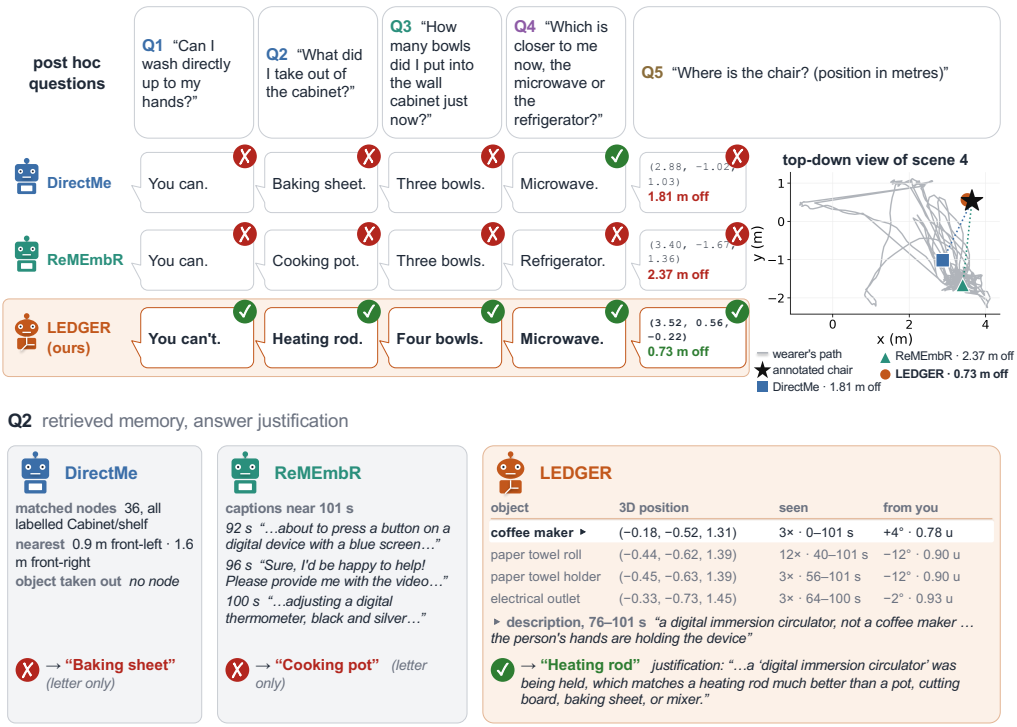


Figure 3: Three memories answer the same questions from their stored records alone, without frames. Q1–Q4 come from one UCS-Bench stream that joins three EPIC-Kitchens recordings; Q5 asks for a chair's position in an Ego4D VQ3D clip, with each memory's answer on the top-down map. **Bottom:** what each memory retrieves for Q2. Questions were chosen to show how the memories differ; Figure 8 (Appendix G.2) adds the video timeline.

Table 2: HD-EPIC accuracy (%) on all 2,400 3D Perception and Object Motion questions, prototype-averaged as in Perrett et al. (2025); chance is 20.0. **yes**: video at answer time; **no**: text alone; ■/■ mark this elsewhere. Published rows use other answerers and protocols, and the 2026 entry is a challenge report (Appendix B.3). LEDGER$_{\text{agent}}$: our memory read by ReMEmbR's querying agent (RQ4). Bold: best per evidence group, except the human reference; 95% intervals in Table 6; other answerers in Figure 4b.

| Method | Video at answer | 3D Perception | Object Motion | Mean |
|---|---|---|---|---|
| LongVA (Zhang et al., 2025b) | yes | 32.9 | 22.7 | 27.8 |
| 2025 challenge, 1st place (Taluzzi et al., 2025) | yes | 42.6 | 30.2 | 36.4 |
| EgoAdapt, 2026 (Chen et al., 2026) | yes | **64.9** | **61.6** | **63.3** |
| Human (sample) (Perrett et al., 2025) | yes | 93.8 | 92.7 | 93.3 |
| *Memories rebuilt on our frames, read by GPT-5.4* | | | | |
| DirectMe, graph only (Wang et al., 2026a) | no | 30.6 | 30.8 | 30.7 |
| ReMEmbR (Anwar et al., 2024) | no | 28.7 | 30.9 | 29.8 |
| OSNOM, adapted (Plizzari et al., 2024) | no | 35.4 | 25.0 | 30.2 |
| AMEGO, adapted (Goletto et al., 2024) | no | 30.4 | 38.9 | 34.7 |
| GPT-5.4, blind | no | 27.2 | 32.1 | 29.7 |
| LEDGER + GPT-5.4 | no | **44.7** | **40.5** | **42.6** |
| LEDGER$_{\text{agent}}$ + GPT-5.4 | no | 37.0 | 32.6 | 34.8 |

**RQ3: Which geometric errors matter for spatial answering?** *Errors in the camera reference frame matter more than errors in object coordinates.* Direction answers depend on the camera's heading, so estimated poses cost fixture-location questions 11 points, or 29 without gravity levelling (Table 13). Refining object coordinates by triangulation adds 2.2 points overall, i.e. less than

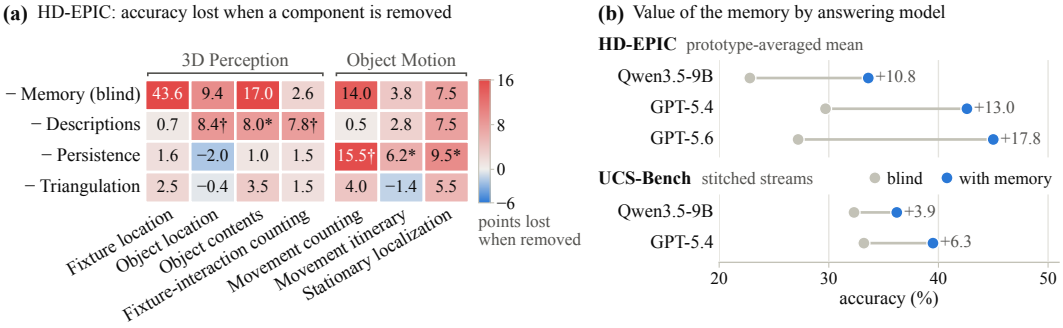


Figure 4: What each component carries, and what the answering model adds. (a) HD-EPIC: accuracy lost on each prototype when one component is removed (2,317 questions on 133 videos; † survives Holm correction, * raw $p < 0.05$; single runs, cells ±7–10 points). (b) Accuracy without and with the memory for each answering model; GPT-5.6 answered with reasoning enabled, the others without. The UCS-Bench component and reader analyses are in Figure 7.

Table 3: UCS-Bench accuracy (%) on 2,771 timestamped questions; 48 frames where frames are given; chance 25.7. DM is DirectMe with its released graph builder and retrieval: "own" adds its graph to the frames on every question, as DirectMe does, and "routed" only on the spatial questions where we add our memory. Bold: best per row among the green and among the red arms. Subcategories and paired tests are in Appendix D.1.

| | | | | **Memory only** | | **Frames + memory** | | |
|---|---|---|---|---|---|---|---|---|
| Category | $n$ | Blind | Frames | DM | LEDGER | DM, own | DM, routed | LEDGER |
| Trajectory & Movement | 648 | 34.9 | 45.4 | 37.3 | **38.6** | 41.8 | 45.5 | **49.1** |
| Position & Orientation | 701 | 23.8 | 39.7 | 26.2 | **33.4** | 37.4 | 37.4 | **43.9** |
| Proximity & Reachability | 538 | 45.4 | **51.7** | 40.0 | **49.3** | 47.2 | 48.5 | 51.3 |
| Category & Quantity | 884 | 33.9 | 42.9 | 21.4 | **36.0** | 35.6 | 42.0 | **43.1** |
| Overall | 2771 | 33.8 | 44.4 | 30.0 | **38.5** | 39.8 | 42.9 | **46.3** |

persistence (5.5) or descriptions (4.9; Table 12). Coordinates are still what make metric localization possible: on VQ3D, our memory has the lowest error of the rebuilt memories on its returned predictions (Table 4). Even there, pose quality bounds the refinement: triangulation improves eligible predictions by about 0.1 m with scan-registered poses but only 0.03 m with feed-forward poses (Table 18).

**RQ4: How much does retrieval affect comparisons between memories?** *Which memory looks better depends on the reader, so comparisons must hold the reader fixed.* We read our record with two readers: our single-shot reader (one call, about 2k input tokens) and ReMEmbR's querying agent (LEDGER$_{\text{agent}}$; three to four calls of 16–23k tokens per question, see Appendix E.7). When the same agent reads both memories, our record beats ReMEmbR's captions by 5.0 points on HD-EPIC and by 26 points in VQ3D queries within 1 m, and is level on UCS-Bench (Tables 2, 4 and 37). ReMEmbR leads natively only on UCS-Bench, 41.4 against our 39.0, and the lead disappears in two steps: event lines on what the wearer did, the weakness found in RQ1, bring our record to 41.0 (Table 28), and the shared agent brings it to 42.1 (Table 37). The agent is not the better reader everywhere: on HD-EPIC our reader grounds the query box in 3D and beats the agent on the same record by 7.8 points (Table 2) at a thirteenth of the input tokens (Table 41).

**RQ5: What breaks as recordings grow longer or span different scenes?** *Longer recordings do not erode the memory's advantage, but scene changes do.* The memory stays 9.9 to 18.0 points above the blind control in every HD-EPIC bin up to 75 minutes, while no rebuilt memory gains more than 5; on UCS-Bench, the 48-frame reader loses accuracy as recordings grow and the memory does not (Figure 5; Appendix F). Across scenes, 100 paired stitched streams show that returning to the same scene costs the memory less than a change of scene (Table 28). The loss arises in both

Table 4: ■VQ3D validation split: memories built by each method on the same 44 clips and read by the same GPT-5.4 matcher from the object's name; errors in metres over graded queries (Appendix C.6). AMEGO stores neither names nor 3D positions. ReMEmbR answers with the wearer's position; OSNOM uses our detections and depth; DirectMe is read without frames. Our rows name the detector and use trajectory memory and EgoLoc poses, without SAM 3 consolidation.

| Memory | Graded / 164 | Median $L_2$ (m) | Mean $L_2$ (m) | $< 1$ m (%) | $< 2$ m (%) |
|---|---|---|---|---|---|
| ReMEmbR (wearer position) | 164 | 1.49 | 1.77 | 23 | 69 |
| DirectMe (graph only) | 164 | 1.91 | 1.98 | 18 | 52 |
| OSNOM (adapted) | 146 | 1.33 | 1.65 | 35 | 71 |
| *No memory: clip centroid* | 164 | 1.69 | 1.82 | 10 | 63 |
| LEDGER, YOLO-World | 148 | 1.06 | 1.46 | 49 | 74 |
| LEDGER, SAM 3 | 153 | **0.99** | **1.41** | **51** | **76** |
| $\text{LEDGER}_{\text{agent}}$, SAM 3 | 164 | 1.00 | 1.54 | 49 | **76** |

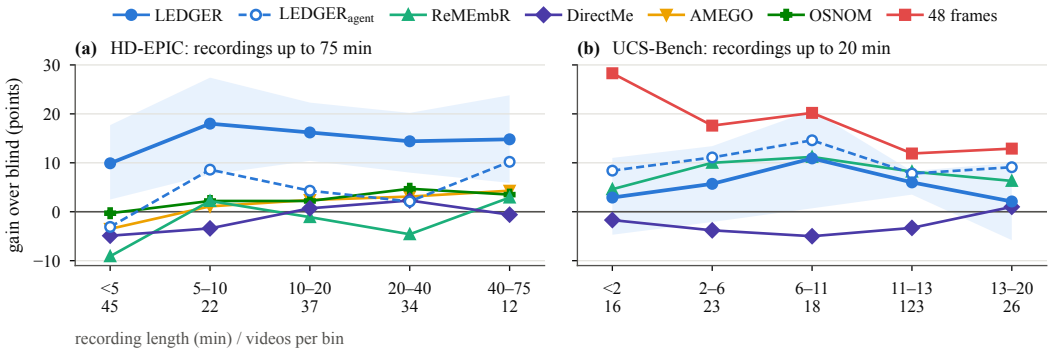


Figure 5: Gain over the blind control by recording length, on the same questions in each bin (question-level accuracy, GPT-5.4). The band is the 95% interval of LEDGER's gain. (a) HD-EPIC, 2,400 questions on 150 videos. (b) UCS-Bench, 3,264 questions on 206 single recordings; AMEGO and OSNOM cannot answer these. Per-bin tables and trend tests are in Appendix F.

stages, because each assumes one scene: retrieval shows records and histories from other scenes, and construction drops labels that cannot belong to one place while its pose gate rejects the whole stream at the seam. Limiting retrieved records and histories to the current scene recovers 3.0 points with true boundaries (Appendix E.3), and per-scene construction recovers half the loss in emulation (Table 30). The fixes are not additive, and these streams also change the camera, wearer, and clip length.

## 5 Conclusion and Limitations

LEDGER makes past object observations reusable for spatial answering and localization. Admitting untouched objects answers fixture questions, and persistence and contextual descriptions contribute more than coordinate refinement; missing event evidence and fragmented identities limit interaction and counting. Comparisons between memories depend on the reader, and scene changes break both retrieval and construction, which per-scene construction partly repairs. Evaluations should measure answerable evidence alongside localization and state retrieval budgets and scene boundaries.

A limitation is that recent frame readers outperform memory-only answering on HD-EPIC, and frames alone outperform memory alone on UCS-Bench, although adding the memory to the frames improves them further. Counting and continuous rest intervals remain unresolved, and construction does not yet scale to the most object-dense hour-long recordings (Appendix D). HD-EPIC uses calibrated poses and some evaluation videos are in our calibration set (Appendix B.9); VQ3D is queried by name rather than crop, and most UCS-Bench memories lack metric scale. Construction is offline, so filtering records by question time does not establish causal online memory.

## AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting dataset and implementing methods. We have not used generative AI tools for proposing or refining hypotheses, designing and providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, interpreting results, and generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting with translation, assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, create or edit software code, edit a research paper to improve readability, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We reviewed LLM generated text to ensure it was consistent with our findings, LLM-generated code was verified and tested for correctness by the main author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

No ethics concerns.

## REPRODUCIBILITY STATEMENT

The full source code of our method is available at an anonymous repository, `https://anonymous.4open.science/r/LEDGER-54FF/README.md`. In addition, the supplementary zip file contains visualizations of sample outputs, namely the memory construction steps and the resulting memories for several long egocentric videos, which can be viewed locally in a browser by opening the included `index.html`.

## REFERENCES


Mohamad Al Mdfaa, Svetlana Lukina, Timur Akhtyamov, Arthur Nigmatzyanov, Dmitrii Nalberskii, Sergey Zagoruyko, and Gonzalo Ferrer. Spatiotemporal knowledge graphs as persistent scene memory for embodied question answering, 2025.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation, 2024.

Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends Cogn. Sci.*, 10(12): 551–557, December 2006.

Yifei Cao, Yu Liu, Guolong Wang, Zhu Liu, Kai Wang, Xianjie Zhang, Jizhe Yu, and Xun Tu. EAGLE: Episodic appearance- and geometry-aware memory for unified 2d-3d visual query localization in egocentric vision. *arXiv preprint arXiv:2511.08007*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Jacob Chalk, Saptarshi Sinha, Dima Damen, Yannis Kalantidis, and Diane Larlus. Whareformer: Learning to track what is where in long egocentric videos. *arXiv preprint arXiv:2607.08537*, 2026.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Guozhi Qiu, Weili Guan, and Liqiang Nie. EgoAdapt: A multi-scene egocentric adaptation method for CVPR 2026 HD-EPIC VQA challenge, 2026. URL https://arxiv.org/abs/2605.24500.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.html.

Tianyi Cheng, Dandan Shan, Ayda Sultan Hassen, Richard Ely Locke Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6ldTxwhgtP.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024b. URL https://arxiv.org/abs/2406.07476.

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.

Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, November 2017.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding, 2024.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6342–6352. IEEE, 2025.

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

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Jason Helbing, Dejan Draschkow, and Melissa L.-H. Võ. Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, 196:104147, 2020. URL https://api.semanticscholar.org/CorpusID:210940013.

Andrew Hollingworth. The relationship between online visual representation of a scene and long-term scene memory. *J. Exp. Psychol. Learn. Mem. Cogn.*, 31(3):396–411, May 2005.

Andrew Hollingworth and John M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28:113–136, 2002. URL https://api.semanticscholar.org/CorpusID:17274346.

Weikai Huang, Jieyu Zhang, Sijun Li, Taoyang Jia, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Matthew Wallingford, Rustin Soraki, Chris Dongjoo Kim, Shuo Liu, Donovan Clay, Taira Anderson, Winson Han, Ali Farhadi, Bharath Hariharan, Zhongzheng Ren, and Ranjay Krishna. Wilddet3d: Scaling promptable 3d detection in the wild, 2026. URL https://arxiv.org/abs/2604.08626.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online, 2025. URL https://arxiv.org/abs/2501.00584.

Almut Hupbach, Oliver Hardt, Rebecca Gomez, and Lynn Nadel. The dynamics of memory: context-dependent updating. *Learn. Mem.*, 15(8):574–579, August 2008.

Savya Khosla, Alexander Schwing, Derek Hoiem, et al. Relocate: A simple training-free baseline for visual query localization using region-based representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3697–3706, 2025.

M Land, N Mennie, and J Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647.

Yiren Lu, Yi Du, Disheng Liu, Yunlai Zhou, Chen Wang, and Yu Yin. GSMem: 3d gaussian splatting as persistent spatial memory for zero-shot embodied exploration and reasoning. *arXiv preprint arXiv:2603.19137*, 2026.

Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. R4DSG: Relative 4D scene graph memory for object-centric question answering in long egocentric video. In *ACM International Conference on Multimedia (MM)*, 2026. arXiv:2608.11017.

Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 45–57. IEEE, 2023.

Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Hybrid structure-from-motion and camera relocalization for enhanced egocentric localization, 2024. 1st place, Ego4D VQ3D challenge 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zaira Manigrasso, Matteo Dunnhofer, Antonino Furnari, Moritz Nottebaum, Antonio Finocchiaro, Davide Marana, Rosario Forte, Giovanni Maria Farinella, and Christian Micheloni. Online episodic memory visual query localization with egocentric streaming object memory. *arXiv preprint arXiv:2411.16934*, 2024.

OpenAI. GPT-5.4 Thinking System Card. OpenAI Deployment Safety Hub, March 2026a. URL https://deploymentsafety.openai.com/gpt-5-4-thinking. Accessed: 2026-09-24.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026b. URL https://openai.com/index/gpt-5-6/.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *International Conference on 3D Vision (3DV)*, 2024. arXiv:2404.05072.

Albert Postma, Roy P C Kessels, and Marieke van Asselen. How the brain remembers and forgets where things are: the neurocognition of object-location memory. *Neurosci. Biobehav. Rev.*, 32(8): 1339–1345, October 2008.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling Video LLMs with Active Real-Time Interaction via Disentangled Perception, Decision, and Reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02239. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02239.

Ivan Rodin, Tz-Ying Wu, Kyle Min, Sharath Nittur Sridhar, Antonino Furnari, Subarna Tripathi, and Giovanni Maria Farinella. EASG-Bench: Video Q&A benchmark with egocentric action scene graphs, 2025. URL https://arxiv.org/abs/2506.05787.

Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments. *arXiv preprint arXiv:2402.13817*, 2024.

You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

Agnese Taluzzi, Davide Gesualdi, Riccardo Santambrogio, Chiara Plizzari, Francesca Palermo, Simone Mentasti, and Matteo Matteucci. From pixels to graphs: using scene and knowledge graphs for hd-epic vqa challenge. *arXiv preprint arXiv:2506.08553*, 2025.

Hao Tang, Kevin Liang, Matt Feiszli, and Weiyao Wang. EgoTracks: A long-term egocentric visual object tracking dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique

Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao,

Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru,

Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

E C Tolman. Cognitive maps in rats and men. *Psychol. Rev.*, 55(4):189–208, July 1948.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

R F Wang and E S Spelke. Updating egocentric representations in human navigation. *Cognition*, 77 (3):215–250, December 2000.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User-centric continual spatial intelligence reasoning in egocentric video streams, 2026a. URL https://arxiv.org/abs/2606.15200.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. EgoMemReason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. *arXiv preprint arXiv:2605.09874*, 2026b.

William H. Warren, Daniel B. Rothman, Benjamin H. Schnapp, and Jonathan D. Ericson. Wormholes in virtual space: From cognitive maps to cognitive graphs. *Cognition*, 166:152–163, 2017. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.05.020. URL https://www.sciencedirect.com/science/article/pii/S0010027717301373.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RnXS7aK4rK.

Yinsong Xu, Wei Jing, Liuxin Zhang, Wanjun Lv, and Hui Li. Semantic and visual evidence for efficient long-video reasoning: A solution for the HD-EPIC VQA challenge, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b.

Sicheng Yang, Yukai Huang, Shitong Sun, Weitong Cai, Jiankang Deng, Jifei Song, and Zhensong Zhang. Optimizing multimodal LLMs for egocentric video understanding: A solution for the HD-EPIC VQA challenge, 2026. 2nd place, HD-EPIC VQA challenge 2025.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Zihao Yu, Xiu Yuan, and Chongjie Zhang. MEMORA: Embodied action memory from egocentric videos for reasoning and planning. *arXiv preprint arXiv:2607.14252*, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=30RAWQVGlx.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025c. ISSN 2835-8856. URL https://openreview.net/forum?id=EElFGvt39K.

Rui Zhou, Xander Yap, Jianwen Cao, Allison Lau, Boyang Sun, and Marc Pollefeys. Memory over maps: 3d object localization without reconstruction. *arXiv preprint arXiv:2603.20530*, 2026.

# APPENDIX CONTENTS


**A Memory Construction Details** **20**
- A.1 Sampling, vocabulary, and detection . . . 20
- A.2 Camera poses and object lifting . . . 20
- A.3 Association and consolidation . . . 20
- A.4 Persistence and contextual descriptions . . . 21
- A.5 Multi-view refinement . . . 21

**B HD-EPIC Details** **21**
- B.1 Evaluation scope and per-prototype results . . . 21
- B.2 Implementation and query adapters . . . 23
- B.3 Published baselines and aggregation . . . 24
- B.4 Memories rebuilt on our setup . . . 25
- B.5 Component ablations . . . 26
- B.6 Design ablations . . . 26
- B.7 Earlier retrieval, oracle, and geometry analyses . . . 27
- B.8 Movement and stationary-state analysis . . . 28
- B.9 Calibration data and cost . . . 28

**C VQ3D Localization and Discovery Analysis** **29**
- C.1 Protocol and interpretation . . . 29
- C.2 Implementation settings . . . 29
- C.3 Localization across memory variants . . . 29
- C.4 Triangulated against monocular positions . . . 29
- C.5 Published methods on the validation split . . . 30
- C.6 Memories built by prior methods on the same clips . . . 31
- C.7 Detector and vocabulary analysis . . . 32

**D UCS-Bench Details** **33**
- D.1 Benchmark, construction, and answering protocol . . . 33
- D.2 Counting . . . 35
- D.3 Calibration set and causal construction . . . 36
- D.4 Implementation settings . . . 36

**E Multi-Scene UCS-Bench Streams** **36**
- E.1 Stream design . . . 36
- E.2 Where the loss appears . . . 37
- E.3 Construction and retrieval . . . 38
- E.4 Scene-cut detection and per-scene construction . . . 38
- E.5 Event lines . . . 39
- E.6 Leave-one-out ablation . . . 39
- E.7 Record against reader . . . 39
- E.8 Open answering models . . . 40
- E.9 Delayed questions . . . 41

**F Accuracy by Recording Length** **41**

**G Failure Analysis and Qualitative Example** **43**
- G.1 When an answer is wrong, where was it lost? . . . 43
- G.2 Three memories on one stream . . . 43

**H Answer-Time Token Cost** **44**


## READING GUIDE

| Main text | Appendices with the details |
|---|---|
| Method (Section 3) | A |
| Benchmarks, baselines, and protocols (Section 4) | B.1–B.4, C.1–C.2, C.5–C.6, D.1, D.4, E.1 |
| RQ1: which questions stored records answer (Section 4) | B.1, B.3–B.4, D.1–D.2, E.8 |
| RQ2: genuine movement or observation noise (Section 4) | B.5, B.8 |
| RQ3: which geometric errors matter (Section 4) | B.5–B.7, C |
| RQ4: retrieval and memory comparisons (Section 4) | E.5, E.7 |
| RQ5: what breaks across scenes (Section 4) | F, E.1–E.4, E.9 |
| Where wrong answers are lost; a worked example (Section 4) | G.1, G.2 |
| Why a memory rather than frames per question (Section 3) | H |

# A Memory Construction Details

## A.1 Sampling, vocabulary, and detection

We sample one frame every four seconds, up to 1,200 frames per recording. Qwen3.5-9B names the visible objects. We merge synonymous names into a vocabulary of at most 80 labels per recording. On UCS-Bench, this step uses at most 80 frames (Appendix D). YOLO-World (Cheng et al., 2024a) then detects objects matching these labels. On VQ3D, where localization accuracy is the measured quantity, we detect with SAM 3 (Carion et al., 2025) instead; YOLO-World is used on the other benchmarks for its lower detection cost (Table 22). Which objects enter memory depends on the vocabulary, the detector's confidence threshold, and the number of detections we can lift into 3D. Dataset-specific resolutions, thresholds, and budgets are listed in Appendices B.2, C, and D.

## A.2 Camera poses and object lifting

We use calibrated device trajectories on HD-EPIC and registered EgoLoc poses on VQ3D. To estimate poses, we use FastVGGT (Shen et al., 2025), a training-free acceleration of VGGT (Wang et al., 2025). It processes 96-frame chunks with 16 overlapping frames. We align adjacent chunks using a Sim(3) transformation estimated from their shared frames.

These estimated poses have arbitrary scale. To recover metric scale, the pipeline compares pairs of observations of the same static object with their predicted metric positions in camera coordinates. This step fails for most UCS-Bench recordings and all TeleEgo recordings. Distances in those memories are therefore expressed in scene units (Appendix D.1).

On UCS-Bench, we first try Depth Anything 3 (Lin et al., 2025), except for circular-fisheye recordings. It processes 48-frame chunks with 12 overlapping frames. We keep its poses only if they pass three consistency checks: the chunk-alignment residual is at most 0.5 m, chunk scales lie within $[0.8, 1.25]$, and the camera moves no more than 6 m between sampled frames. Otherwise, we fall back to FastVGGT. If FastVGGT cannot align a chunk with the preceding one, it starts from a new anchor (Appendix D).

Given an image, a 2D box, and an object label, WildDet3D (Huang et al., 2026) predicts a metric 3D box in camera coordinates. We use its centre as the object's position. This avoids selecting pixels from a dense depth map, where object and background surfaces can be mixed. Although the predicted box is metric, its position in scene coordinates is only metric if the camera poses also have a known scale.

**Geometry-estimator analysis.** In the earlier HD-EPIC analysis, the median localization error was 0.39 m, with a median along-ray component of 0.37 m (Appendix B.7). This motivates using multiple views with known camera poses to refine object positions.

On the UCS-Bench calibration set, two Depth Anything 3 reconstructions were rejected. Their chunk scale ratios were 1.16 and 1.58, and their alignment residuals at chunk boundaries were 0.67 and 7.10 m, respectively. We fall back to FastVGGT, which also handles fisheye recordings but does not provide metric scale on its own.

We compare both pose estimators on the same frames from all 44 VQ3D clips (Appendix C). After aligning each trajectory to EgoLoc camera centres with one Sim(3) transformation per clip, their median position errors are similar: 9.1 cm for Depth Anything 3 and 9.2 cm for FastVGGT. Median orientation errors are 5.5° and 5.9°, respectively. However, Depth Anything 3 fails to join its chunks into a single trajectory on 11 clips. Object localization also shows no clear difference: median paired differences range from $-0.03$ to $+0.02$ m across detectors and memory variants, with every interval including zero. These results support using FastVGGT as the fallback when poses are unavailable, although both estimators still require metric scale recovery.

## A.3 Association and consolidation

Trajectory memory clusters observations in 3D, associating them chronologically by label and proximity to an instance's most recent position. Detections in the same frame remain separate unless they are overlapping duplicates. A second pass, called consolidation, reconnects tracks that may belong

to the same object. It groups tracks with synonymous labels, uses the SAM 3 video tracker (Carion et al., 2025) to propagate their masks across frames, and merges tracks whose propagated masks agree. Co-occurrence and 3D-distance checks block incompatible merges.

We use consolidation on HD-EPIC and UCS-Bench. On VQ3D, we compare raw, trajectory, and consolidated memories, using a trajectory association radius of 1.0 m and a consolidation mask-IoU threshold of 0.3. On UCS-Bench, merges are blocked beyond a 3D distance of 1.0, measured in scene units when metric scale is unavailable (Appendix D.2). Even with consolidation, an object may receive a new identity after a long absence or a large move.

### A.4 Persistence and contextual descriptions

Distance-threshold segmentation starts a new segment as soon as an observation lies more than $\tau$ from the current resting position. Our persistence rule waits until $k$ successive observations of the object lie outside this radius before confirming a new segment. If observations return within the radius before confirmation, they remain in the current segment.

On HD-EPIC, we use $\tau = 0.3$ m and $k = 3$. We use the same values on UCS-Bench, but measure the radius in scene units when metric scale is unavailable. Appendix B.8 compares the two rules.

For up to six segments per object, Qwen3.5-9B describes a representative frame with the object outlined. These descriptions record details such as the object's contents, its supporting surface, and interactions that positions alone do not capture.

In the reported HD-EPIC runs, we first generated descriptions for distance-threshold segments. Each persistence segment then inherited the description of the segment with the greatest temporal overlap. The persistence ablation therefore also changes which descriptions appear in the records; it does not keep all textual evidence fixed (Appendix B.2). The additional activity descriptions for sampled frames in the UCS-Bench extensions are detailed in Appendix E.5.

### A.5 Multi-view refinement

Observing a resting object from several viewpoints helps constrain its 3D position. Each detected box centre defines a viewing ray from the camera. Given camera centres $\mathbf{c}_i$ and unit ray directions $\mathbf{d}_i$, we estimate the point that minimizes the sum of squared distances to these rays:

$$\hat{\mathbf{x}} = \left( \sum_i \mathbf{P}_i \right)^{-1} \sum_i \mathbf{P}_i \mathbf{c}_i, \qquad \mathbf{P}_i = \mathbf{I} - \mathbf{d}_i \mathbf{d}_i^\top . \tag{1}$$

We accept this estimate only when the camera positions are sufficiently separated, the ray geometry supports a stable solution, and the estimate stays close to the monocular prediction. On HD-EPIC, we require at least three rays, a camera baseline of at least 0.10 m, a minimum eigenvalue of 0.25 for $\sum_i \mathbf{P}_i$, and a shift of no more than 2.0 m from the monocular prediction (Table 7). The localization analysis and question-answering ablation are reported in Appendices B.7 and B.5.

## B HD-EPIC Details

### B.1 Evaluation scope and per-prototype results

HD-EPIC contains 26,550 five-way multiple-choice questions across 30 prototypes and seven categories. We evaluate all questions in 3D Perception and Object Motion, covering 150 videos (Table 5; Figure 6). These questions have no official training, validation, or test split.

We build memories using the released camera poses and reserve ground-truth geometry for analysis. Across the 150 videos (40.7 hours), 36,633 sampled frames produce 52,001 object instances, 504,981 observations, and 82,445 rest segments. Answering from the same memory in a separate session yields 41.9% accuracy, compared with 42.6% originally, illustrating a 0.7-point variation between these two runs.

Component ablations use 2,317 questions from the 133 videos processed in the initial pass, with the same prototype-averaged metric. The earlier ten-video analyses reported below use the calibration set and different run snapshots.

Table 5: HD-EPIC accuracy (%) on all seven evaluated prototypes. Memory scores are bold when they exceed the matched blind control.

| Category | Prototype | $n$ | Qwen3.5-9B Blind | Qwen3.5-9B Memory | GPT-5.4 Blind | GPT-5.4 Memory |
|---|---|---|---|---|---|---|
| 3D Perception | object location | 500 | 25.4 | **45.6** | 38.6 | **48.0** |
| | contents retrieval | 200 | 16.0 | **28.5** | 16.5 | **33.5** |
| | fixture location | 500 | 16.2 | **52.6** | 13.8 | **56.0** |
| | fixture interaction counting | 300 | 36.0 | **40.3** | 40.0 | **41.3** |
| Object Motion | movement itinerary | 500 | 11.8 | **26.2** | 34.8 | **38.6** |
| | movement counting | 200 | 28.5 | **35.5** | 31.0 | **45.0** |
| | stationary localization | 200 | 26.0 | 14.5 | 30.5 | **38.0** |
| Overall (question average) | | 2,400 | 21.5 | **37.5** | 29.7 | **44.6** |

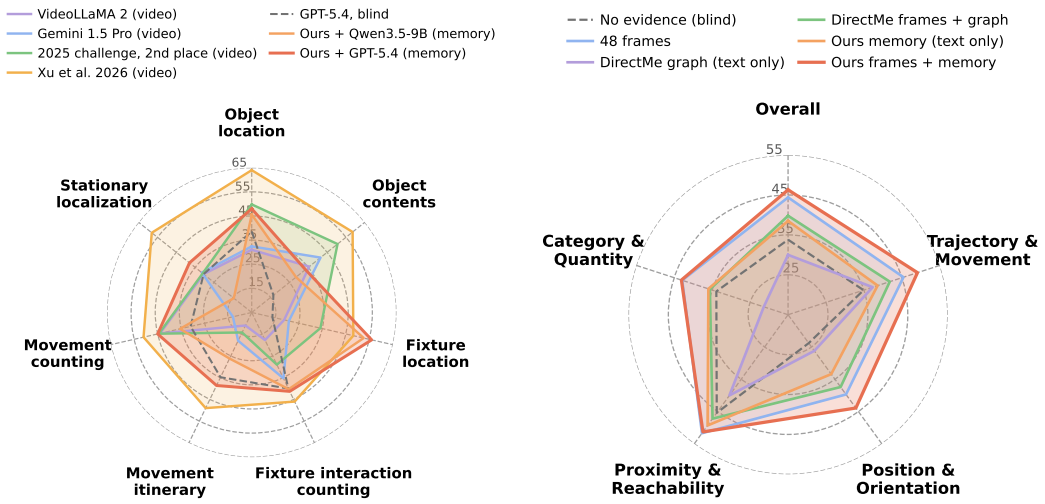


Figure 6: Accuracy by question type. **Left:** HD-EPIC by prototype; published video models as printed in Xu et al. (2026) (EgoAdapt reports no per-prototype scores). **Right:** UCS-Bench by category on the 2,771 questions of Table 3.

Table 6: Prototype-averaged HD-EPIC accuracy (%) with 95% bootstrap intervals (2,000 resamples stratified by prototype, seed 0). Memory improves mean accuracy over the matched blind control by 10.8 points $[8.2, 13.3]$ for Qwen3.5-9B and 13.0 points $[10.4, 15.5]$ for GPT-5.4.

| Answerer | Evidence | 3D Perception | Object Motion | Mean |
|---|---|---|---|---|
| Qwen3.5-9B | Blind | 23.4 [21.2,25.7] | 22.1 [19.0,25.1] | 22.8 [20.9,24.7] |
| | Memory | **41.8** [39.3,44.4] | **25.4** [22.3,28.4] | **33.6** [31.6,35.5] |
| GPT-5.4 | Blind | 27.2 [25.0,29.6] | 32.1 [28.9,35.4] | 29.7 [27.7,31.7] |
| | Memory | **44.7** [42.0,47.3] | **40.5** [37.2,44.3] | **42.6** [40.5,44.8] |

We report 95% bootstrap intervals using 2,000 resamples stratified by prototype (seed 0). Comparisons use the same questions in both conditions.

We also report pooled accuracy, which gives each question equal weight, so larger prototypes contribute more. For Qwen, memory raises pooled 3D Perception accuracy from 23.2% to 44.6% and Object Motion accuracy from 18.7% to 25.7%. For GPT, the corresponding improvements are from 27.7% to 47.4% and from 33.0% to 39.9%.

Across all 2,400 questions, pooled accuracy rises from 21.5% to 37.5% for Qwen and from 29.7% to 44.6% for GPT. The paired gains are 16.0 points $[13.6, 18.3]$ and 14.9 points $[12.7, 17.3]$, respectively. These pooled scores should not be compared directly with the HD-EPIC paper, which averages prototype scores.

The intervals describe uncertainty for the evaluated videos. They do not account for clustering of questions within videos or replace evaluation on held-out videos.

### B.2 Implementation and query adapters

We convert the Aria fisheye RGB frames into $1802 \times 1802$ virtual pinhole images. We combine the released device poses with the camera calibration to obtain the world-to-camera transformation: $\mathbf{T}_{cw} = (\mathbf{T}_{wd}\mathbf{T}_{dc})^{-1}$. To check the coordinate conventions, we project 35,722 ground-truth object locations into the images. Of these, 98.3% fall inside their annotated boxes, with a median offset of 22 pixels. These annotations are not used to build the memory.

Table 7: HD-EPIC construction and answering settings.

| Stage | Parameter | Value |
|---|---|---|
| Sampling | interval / cap | 4 s / 1,200 frames |
| Rectification | pinhole resolution | $1802 \times 1802$ px |
| Tagging | model / token cap / vocabulary cap | Qwen3.5-9B / 130 / 80 |
| Detection | model | YOLO-World v8x-worldv2 |
| | input / confidence / top-$k$ | 1408 px / 0.05 / 40 |
| 3D lifting | per-frame / per-video / per-tag budget | 12 / $\geq 400$ / $\max(25, N/4)$ |
| Triangulation | rays / baseline / $\lambda_{\min}$ / shift | 3 / 0.10 m / 0.25 / 2.0 m |
| Association | tracker IoU / 3D merge veto | 0.3 / 1.0 m |
| Segmentation | radius $\tau$ / persistence $k$ | 0.3 m / 3 |
| Descriptions | frame size / segments per object | 896 px / 6 |
| Box retrieval | time window / candidates | $\pm 6$ s / 4 |
| Answering | protocol | one call, options A–E |

For questions that specify an object with a box, we use the sampled frame nearest the query time to estimate the box's 3D position. We rank object instances observed near that time by their minimum distance to this position and retrieve the closest four. We then replace the box reference in the question with the estimated coordinate. The answerer receives object labels, ranking signals, and segment histories as text. It sees no images, although grounding the query box still uses a frame.

For fixture-location questions, we match names using predefined synonym groups, such as hob/stove/cooktop and fridge/refrigerator/freezer. We retrieve the four matching instances with the most observations. To compute their clock directions, we project the camera's forward axis and the direction to each object onto the horizontal plane: 12 is ahead, 3 is right, 6 is behind, and 9 is left.

Using ground-truth fixture positions, this calculation selects the correct direction option on 97.8% of 365 validation questions. With the constructed memory, retrieval finds a name-matched fixture for 83% of the 500 fixture-location questions. Choosing the option closest to the first candidate's clock direction scores 59.2% across all 500 questions, compared with 56.0% for GPT reading the candidate histories. This geometric rule applies specifically to fixture-direction questions.

We run Qwen with greedy decoding and thinking disabled. We run GPT-5.4 with reasoning effort set to none and temperature 1.0 in all experiments, so its answers can vary across otherwise identical runs. Mean usage per question is 1,938 input and 381 output tokens for Qwen, and 1,750 input and 231 output tokens for GPT.

Descriptions are generated for distance-threshold segments. Each persistence segment inherits the description of the segment with the greatest temporal overlap. The segmentation ablation therefore changes both segment boundaries and the assignment of descriptions; it does not keep all textual evidence fixed.

### B.3 PUBLISHED BASELINES AND AGGREGATION

The HD-EPIC paper averages prototype accuracies to obtain category scores (Perrett et al., 2025). We follow this convention for our primary results. We also report pooled accuracy, which gives each question equal weight. For category $C$, with prototype accuracies $a_p$ and question counts $n_p$, the two scores are

$$\text{Prototype average} = \frac{1}{|C|}\sum_{p\in C} a_p, \qquad \text{Pooled accuracy} = \frac{\sum_{p\in C} n_p a_p}{\sum_{p\in C} n_p}.$$

Our rows in Table 8 use prototype averages to match the published convention. The number of correct answers for each prototype can be uniquely recovered from its rounded accuracy and question count in Table 5.

Table 8: Published HD-EPIC category scores and our prototype-averaged scores (%). Published baselines were not rerun; models, visual inputs, and inference budgets differ. The team labels used in the challenge reports are inconsistent between papers, so rows are identified by paper: the 40.9 / 29.9 entry is the method of Yang et al. (2026) (2nd place 2025), and the 42.6 / 30.2 entry is the scene- and knowledge-graph system of Taluzzi et al. (2025) (1st place 2025, 44.21 overall).

| Model | Images to answerer | 3D Perception | Object Motion |
|---|---|---|---|
| Llama 3.2, text only (Perrett et al., 2025) | **no** | 22.3 | 25.5 |
| Gemini 1.5 Pro, text only (Perrett et al., 2025) | **no** | 21.5 | 27.7 |
| VideoLLaMA 2 (Cheng et al., 2024b) | **yes** | 25.7 | 28.5 |
| LongVA (Zhang et al., 2025b) | **yes** | 32.9 | 22.7 |
| LLaVA-Video (Zhang et al., 2025c) | **yes** | 27.3 | 18.9 |
| Gemini 1.5 Pro (Team et al., 2024) | **yes** | 32.5 | 20.8 |
| Qwen2.5-VL-7B, zero-shot (Yang et al., 2026) | **yes** | 35.0 | 23.9 |
| Qwen2.5-VL-32B, zero-shot (Yang et al., 2026) | **yes** | 35.6 | 19.8 |
| 2025 challenge, 2nd place (Yang et al., 2026) | **yes** | 40.9 | 29.9 |
| 2025 challenge, 1st place (Taluzzi et al., 2025) | **yes** | 42.6 | 30.2 |
| Gemini 3.1 Pro + object detector, 2026 (Xu et al., 2026) | **yes** | 54.2 | 52.7 |
| EgoAdapt, 2026 (Chen et al., 2026) | **yes** | 64.9 | 61.6 |
| Human (sample) (Perrett et al., 2025) | **yes** | 93.8 | 92.7 |
| Qwen3.5-9B blind (ours) | no | 23.4 | 22.1 |
| LEDGER + Qwen3.5-9B | no | 41.8 | 25.4 |
| GPT-5.4 blind (ours) | no | 27.2 | 32.1 |
| LEDGER + GPT-5.4 | no | 44.7 | 40.5 |

Table 9: HD-EPIC accuracy per prototype (%) for the published entries that report it (as printed in Xu et al. (2026)) and for our memory. Bold marks the best score per prototype. EgoAdapt reports category means only.

| Prototype | $n$ | Video-LLaMA 2 | LLaVA-Video | Gemini 1.5 Pro | 2025 2nd | Xu et al. 2026 | GPT blind | Ours + Qwen | Ours + GPT |
|---|---|---|---|---|---|---|---|---|---|
| Fixture interaction counting | 300 | 17.7 | 16.3 | 35.3 | 29.0 | **46.0** | 40.0 | 40.3 | 41.3 |
| Fixture location | 500 | 18.8 | 21.8 | 20.8 | 34.2 | 48.2 | 13.8 | 52.6 | **56.0** |
| Object location | 500 | 31.0 | 30.6 | 32.4 | 49.8 | **64.0** | 38.6 | 45.6 | 48.0 |
| Object contents retrieval | 200 | 35.5 | 40.5 | 41.5 | 50.5 | **58.5** | 16.5 | 28.5 | 33.5 |
| Movement itinerary | 500 | 11.0 | 9.8 | 18.0 | 14.2 | **49.0** | 34.8 | 26.2 | 38.6 |
| Movement counting | 200 | 44.0 | 20.0 | 13.0 | 44.5 | **51.0** | 31.0 | 35.5 | 45.0 |
| Stationary object localization | 200 | 30.5 | 27.0 | 31.5 | 31.0 | **58.0** | 30.5 | 14.5 | 38.0 |

Using GPT, the memory exceeds the video baselines of the HD-EPIC paper and both 2025 challenge entries under this averaging convention, and it remains below the two 2026 challenge reports, which read the video with newer models. EgoAdapt uses category-specific video sampling and option-order voting. Neither comparison isolates the benefit of memory; our matched blind runs provide that control. We also tested a newer model (GPT-5.6-sol) on the calibration set of ten videos (418 questions), using the same answering protocol with three evidence conditions: 50 frames, memory,

or no evidence. Prototype-averaged 3D Perception and Object Motion scores were 57.9% and 45.8% with frames, 40.1% and 33.4% with memory, and 23.6% and 32.4% for the blind control. Frames outperform memory in this test, consistent with the published 2026 results.

The per-prototype results reveal where memory helps most (Table 9). Fixture location is the only prototype where our memory exceeds every reported baseline score. Movement counting and fixture-interaction counting are comparable to the 2025 entries. The weakest results are on object contents retrieval and, for Qwen, stationary object localization.

### B.4 Memories rebuilt on our setup

Four memory-based methods are rebuilt on the same 150 videos and 2,400 questions. Each method uses its released code to build memory from the same rectified frames, sampled every 4 s. GPT-5.4 answers from each memory without seeing the video, and all scores use the same prototype averaging. For questions that identify an object with a box, we match the query box to each memory's own observations at the query time. We select up to three matches with the highest 2D overlap, requiring IoU $\geq 0.1$. Method-specific adaptations:

- **DirectMe** (Wang et al., 2026a): released graph builder (Objects365 vocabulary, YOLO-World, SAM 2, Depth Anything 3 depth and poses) and its text-only graph prompt without keyframes, evaluated over the whole video, with the pose for egocentric relations taken at the question's time. Its object-count shortcut is disabled because HD-EPIC counting questions count events, not instances.
- **ReMEmbR** (Anwar et al., 2024): released VILA captioner (one caption per sampled frame), camera positions from the released device trajectories, and its retrieval agent with the options appended; memory times are video times.
- **OSNOM** (Plizzari et al., 2024): released tracker with its published configuration, given the detections our memory lifts (as box masks), the released device trajectories, and FastVGGT depth scaled to metres per 120-frame chunk in place of its mesh-rendered depth. It has no question-answering stage; a common harness serializes the matched track's position history, nearby tracks, and, for fixture questions, name-matched tracks with clock directions computed as in our agent.
- **AMEGO** (Goletto et al., 2024): its hand–object detector weights are no longer available, so the Hands23 detector (Cheng et al., 2023) from the same group replaces it. Its frame-count thresholds, set for 50–60 fps, give an empty memory at 0.25 fps and are reduced to one frame; flow filtering and the object tracker are disabled. AMEGO stores no object names and no 3D positions.

The 95% bootstrap intervals on the mean are $[28.8, 32.8]$ for DirectMe, $[27.8, 31.8]$ for ReMEmbR, $[28.2, 32.2]$ for OSNOM, and $[32.6, 36.7]$ for AMEGO, against $[40.5, 44.8]$ for our memory. Read by the same ReMEmbR querying agent, our memory exceeds ReMEmbR's captions by 5.0 points $[+2.8, +7.2]$ (exact McNemar $p = 1.0 \times 10^{-9}$; Table 2).

Table 10: HD-EPIC accuracy per prototype (%, rounded) for the memories rebuilt on our setup. Bold marks the best memory per prototype.

| Prototype | $n$ | Blind | DirectMe | ReMEmbR | OSNOM | AMEGO | Ours |
|---|---|---|---|---|---|---|---|
| Fixture location | 500 | 14 | 26 | 21 | 44 | 13 | **56** |
| Object location | 500 | 39 | 37 | 39 | 42 | 40 | **48** |
| Object contents retrieval | 200 | 16 | 20 | 30 | 21 | 22 | **34** |
| Fixture interaction counting | 300 | 40 | 39 | 25 | 35 | **47** | 41 |
| Movement itinerary | 500 | 35 | 16 | 14 | 19 | 27 | **39** |
| Movement counting | 200 | 31 | 37 | 32 | 16 | 34 | **45** |
| Stationary object localization | 200 | 30 | 39 | 47 | 40 | **56** | 38 |

### B.5 COMPONENT ABLATIONS

"Segment descriptions" removes all descriptions and summaries from the final memory; "persistence segmentation" uses the distance-threshold rule ($\tau = 0.3$ m, no persistence); "multi-view triangulation" uses monocular depth with the same persistence segmentation. Table 12 reports category-level effects and their intervals. Per-prototype comparisons use the same questions in both conditions (Table 11). We use exact McNemar tests, with Holm correction across the 21 prototype–component comparisons, and bootstrap intervals from 4,000 resamples of videos.

Three effects remain significant after correction: descriptions improve object location by 8.4 points $[3.0, 13.4]$ and fixture-interaction counting by 7.8 points $[3.8, 12.0]$; persistence improves movement counting by 15.5 points $[8.1, 23.2]$.

Each condition uses a single GPT-5.4 run. Repeating identical prompts changes whether 15–20% of answers are correct, and individual per-prototype scores vary by about $\pm$7–10 points.

Table 11: Accuracy lost on each HD-EPIC prototype when one component is removed from the memory (full − ablated, points; 2,317 questions from 133 videos). $^{\dagger}$Significant after Holm correction over the 21 tests; $^{*}$raw $p < 0.05$. One run per arm, so single cells vary by about $\pm$7–10 points.

| Prototype | $n$ | Full | Blind | −Descriptions | −Persistence | −Triangulation |
|---|---|---|---|---|---|---|
| *3D Perception* | | | | | | |
| Fixture location | 447 | 56.4 | 12.8 | $+0.7$ | $+1.6$ | $+2.5$ |
| Object location | 500 | 48.0 | 38.6 | $+8.4^{\dagger}$ | $-2.0$ | $-0.4$ |
| Object contents retrieval | 200 | 33.5 | 16.5 | $+8.0^{*}$ | $+1.0$ | $+3.5$ |
| Fixture interaction counting | 270 | 42.2 | 39.6 | $+7.8^{\dagger}$ | $+1.5$ | $+1.5$ |
| *Object Motion* | | | | | | |
| Movement counting | 200 | 45.0 | 31.0 | $+0.5$ | $+15.5^{\dagger}$ | $+4.0$ |
| Movement itinerary | 500 | 38.6 | 34.8 | $+2.8$ | $+6.2^{*}$ | $-1.4$ |
| Stationary localization | 200 | 38.0 | 30.5 | $+7.5$ | $+9.5^{*}$ | $+5.5$ |
| 3D Perception (prototype mean) | 1,417 | 45.0 | 26.9 | $+6.2$ | $+0.5$ | $+1.8$ |
| Object Motion (prototype mean) | 900 | 40.5 | 32.1 | $+3.6$ | $+10.4$ | $+2.7$ |
| Mean | 2,317 | 42.8 | 29.5 | $+4.9$ | $+5.5$ | $+2.2$ |

Table 12: HD-EPIC component ablations at the category level (questions and runs of Table 11; prototype-averaged mean of the two categories). Differences are paired, computed before rounding, with 95% stratified bootstrap intervals.

| Component removed | Without | With | Gain (points) |
|---|---|---|---|
| Persistence segmentation | 37.3 | 42.8 | $+5.5$ [+3.1,+7.9] |
| Segment descriptions | 37.9 | 42.8 | $+4.9$ [+2.9,+6.9] |
| Multi-view triangulation | 40.5 | 42.8 | $+2.2$ [+0.2,+4.4] |

### B.6 DESIGN ABLATIONS

Each row of Table 13 changes one part of the final HD-EPIC system and re-answers all 2,400 questions with the same prompts. We test differences in mean accuracy using paired sign-flip permutation tests with 20,000 draws and Holm correction across the twelve design comparisons. For individual prototypes, we use exact McNemar tests with Holm correction across 84 cells. Only the estimated poses without gravity levelling significantly change the mean after correction. The GPT-5.6, text-retrieval, and 2D-area-retrieval differences are significant only before correction.

The pose-free variants use FastVGGT poses. We rescale their camera centres by $1/1.70$, a factor measured against ground-truth depth on HD-EPIC. These variants therefore avoid the released camera poses but still use ground-truth information for scale. Gravity levelling estimates the up direction as the direction most nearly perpendicular to all camera $x$-axes, with a median error of 6.0°.

The GPT-5.6 blind control includes six API errors and seven unparsed answers, all scored as incorrect. These affect its score by at most 0.3 points. GPT-5.6 (OpenAI, 2026b) uses reasoning on

98.9% of questions, with a median of 366 reasoning tokens. In contrast, GPT-5.4 uses no reasoning tokens. The 2.4-point gain therefore combines a change of model with the addition of reasoning.

The bottom block reports full memory rebuilds on the calibration set of ten videos (418 questions). Each is compared with the final memory re-answered in the same session. Intervals are approximately $\pm 5$ points, and no difference is significant. Sampling every two seconds mainly lowers movement-counting accuracy, although the mean change stays within these intervals. Across the 981 annotated object names in these videos, semantic recall is 61% for Qwen3.5-9B, 60% for GPT-5.4 with vision, and 54% for Gemma-3-4B.

The component removals in Table 11 come from a separate experiment on 2,317 questions with its own full-memory baseline. They are not repeated here.

Table 13: HD-EPIC design ablations: one change from the final system at a time, all 2,400 questions, prototype-averaged accuracy. $\Delta$ is against the final system; the prototype columns give the change on each prototype. Markers and single-run noise as in Table 11, with Holm correction over this table's tests. [§]See the text for the pose rescaling.

| | | | 3D Perception | | | | Object Motion | | |
|---|---|---|---|---|---|---|---|---|---|
| Change from the final system | Mean | $\Delta$ | obj. loc. | contents | fixt. loc. | fixt. count | itinerary | mov. count | stationary |
| Final system (GPT-5.4) | 42.6 | – | 48.0 | 33.5 | 56.0 | 41.3 | 38.6 | 45.0 | 38.0 |
| Answering model: GPT-5.6 | 45.0 | $+2.4^{*}$ | $+0.4$ | $+7.0^{*}$ | $+2.2$ | $+0.7$ | $+5.2^{*}$ | $+0.5$ | $+1.0$ |
| no memory, GPT-5.4 | 29.7 | $-13.0^{*}$ | $-9.4^{*}$ | $-17.0^{*}$ | $-42.2^{*}$ | $-1.3$ | $-3.8$ | $-14.0^{*}$ | $-7.5$ |
| no memory, GPT-5.6 | 27.2 | $-15.4^{*}$ | $-20.0^{*}$ | $-14.5^{*}$ | $-36.8^{*}$ | $-1.0$ | $-21.0^{*}$ | $-25.5^{*}$ | $+8.5$ |
| Lift depth: FastVGGT | 44.7 | $+2.1$ | $+2.2$ | $-5.0$ | $+5.2^{*}$ | $-1.7$ | $+7.8^{*}$ | $+1.5$ | $+2.5$ |
| WildDet3D, depth $\div$ 1.70 | 43.0 | $+0.4$ | $+1.8$ | $-3.5$ | $+2.6$ | $+0.3$ | $+9.0^{\dagger}$ | $-1.5$ | $-6.0$ |
| Depth Anything 3 | 42.0 | $-0.7$ | $+1.0$ | $-4.5$ | $+1.6$ | $-1.0$ | $+6.2^{*}$ | $-0.5$ | $-7.5$ |
| Poses: FastVGGT, gravity-levelled[§] | 40.8 | $-1.9$ | $+4.4^{*}$ | $+0.5$ | $-11.4^{\dagger}$ | $-2.3$ | $+1.0$ | $+2.0$ | $-7.5$ |
| same, not levelled[§] | 39.1 | $-3.5^{\dagger}$ | $+3.8^{*}$ | $+2.0$ | $-29.4^{\dagger}$ | $-2.0$ | $+1.0$ | $+3.0$ | $-6.0$ |
| Retrieval: 2D box IoU | 42.0 | $-0.6$ | $+2.6$ | $+1.0$ | $+0.4$ | $-2.7$ | $+0.8$ | $-3.0$ | $-2.5$ |
| 3D box IoU | 41.0 | $-1.6$ | $-0.6$ | $-0.5$ | $-1.2$ | $+0.0$ | $+1.2$ | $-2.0$ | $-7.0$ |
| 2D area within 1 m | 40.4 | $-2.2^{*}$ | $-1.8$ | $+0.5$ | $-0.6$ | $+0.0$ | $+0.6$ | $-5.5$ | $-7.0$ |
| text only | 39.6 | $-3.0^{*}$ | $+1.8$ | $-8.5^{*}$ | $+0.0$ | $-2.3$ | $-5.2$ | $+0.5$ | $-6.5$ |
| Question rephrasing off | 41.6 | $-1.1$ | $+0.8$ | $-1.0$ | $+0.8$ | $-3.3$ | $-0.8$ | $-1.0$ | $-2.5$ |
| *Whole-memory rebuilds on the calibration set (418 questions; reference: final system re-answered, 33.4)* | | | | | | | | | |
| Calibration reference (final system) | 33.4 | – | 53.2 | 16.7 | 34.0 | 34.4 | 39.3 | 33.3 | 24.3 |
| Detector: SAM 3 | 36.3 | $+2.9$ | $-3.2$ | $+12.5$ | $+12.8$ | $+3.1$ | $+1.2$ | $-2.7$ | $+0.0$ |
| Tagger: Gemma-3-4B | 32.8 | $-0.6$ | $-5.1$ | $+4.1$ | $+0.0$ | $-3.2$ | $+4.7$ | $+0.0$ | $-5.4$ |
| GPT-5.4 vision | 34.9 | $+1.5$ | $-6.4$ | $+0.0$ | $+0.0$ | $+6.2$ | $+3.6$ | $+0.0$ | $+5.4$ |
| Sampling: one frame per 2 s | 28.8 | $-4.7$ | $-5.1$ | $-4.2$ | $+2.2$ | $+0.0$ | $-6.0$ | $-16.6$ | $+0.0$ |

### B.7 Earlier retrieval, oracle, and geometry analyses

The following analyses use an earlier calibration snapshot of ten videos with 339 object-level questions. These results come from separate runs, rather than the current 2,400-question evaluation or the 2,317-question component ablations. Unless stated otherwise, the configurations in Table 14 use distance-threshold segmentation and monocular positions. With GPT, retrieval criteria differ by at most 0.3 accuracy points. Switching the answerer changes accuracy by 7.4 points when using the same 3D-centre retrieval.

Table 14: Earlier 339-question retrieval analysis (accuracy, %).

| Retrieval | Qwen3.5-9B | GPT-5.4 |
|---|---|---|
| Blind control | 19.5 | – |
| 3D centre distance | 29.2 | 36.6 |
| 2D area among candidates within 1 m | 33.9 | 36.6 |
| Union of both sets | 30.4 | 36.3 |
| 3D centre, triangulated positions | – | 40.4 |

The trajectory oracle (Table 15) corrects target selection and history as well as localization. It does not separately identify each failure source. The near-identical retrieved and non-retrieved "any of

Table 15: Earlier 339-question oracle analysis. Ground-truth supporting-fixture labels add information beyond correcting coordinates.

| Evidence | Accuracy (%) |
|---|---|
| Predicted memory, triangulated positions | 40.4 |
| Ground-truth queried-object trajectory, positions only | 52.2 |
| Ground-truth trajectory and supporting-fixture names | 77.6 |
| Any of four single-candidate answers, retrieved objects | 65.5 |
| Any of four single-candidate answers, non-retrieved objects | 64.9 |

four" scores also show why best-of-several answering success cannot be interpreted as retrieved evidence coverage: multiple guesses can produce a high score without selecting relevant objects.

We evaluate triangulation on 389 detections matched to ground-truth boxes with IoU $\geq 0.5$. It reduces median 3D error from 0.39 to 0.35 m and increases the share within 0.25 m of the true position from 13.6% to 37.3% (Equation 1). Before refinement, most error lies along the viewing ray: the median along-ray component is 0.37 m, compared with 0.03 m perpendicular to it.

For 377 detections observed from at least two known camera poses, we group results into quartiles by camera baseline. The share within 0.25 m improves in every group: from 10.5% to 27.4%, 12.8% to 46.8%, 12.7% to 39.2%, and 19.1% to 38.3%. This analysis measures localization of matched detections, not how many objects the system discovers.

### B.8 Movement and stationary-state analysis

Table 16: Earlier movement-count analysis on 89 questions whose queried object is in memory (ground-truth mean count: 2.5).

| Rule | Mean count | Exact | Within $\pm 1$ | Bias |
|---|---|---|---|---|
| Threshold, $\tau = 0.3$ m | 4.9 | 13.5% | 36.0% | +2.3 |
| Threshold, $\tau = 0.8$ m | 2.3 | 22.5% | 57.3% | $-0.3$ |
| Persistence, $\tau = 0.3$ m, $k = 3$ | 2.2 | **28.1%** | **67.4%** | $-0.3$ |

The threshold rule over-counts movement in 64% of these questions at 0.3 m (Table 16). A larger distance threshold suppresses localization noise but can miss small movements. In the annotations, 19% of moves are shorter than 0.3 m and 50% are shorter than 0.8 m. Meanwhile, the 95th percentile of localization error is 0.84 m. Persistence improves movement counts on questions whose target is already in memory, separating segmentation errors from failures to retain the target. Bias and mean counts are rounded separately.

Stationary-localization questions ask when an object stopped moving and remained still. In the earlier analysis, choosing the answer timestamp closest to a rest-segment start is correct for only 22% of questions, compared with 20% chance. This suggests that sampled observation times do not reliably identify when continuous rest began. Memory does not always hurt this question type, however: the current Qwen and GPT runs show opposite effects (Table 5).

### B.9 Calibration data and cost

We developed the segmentation and retrieval rules, triangulation thresholds, and fixture synonym groups on a small calibration set of 5–10 videos kept separate from evaluation.

In an earlier measurement on a 24 GB GPU, WildDet3D took 0.85, 0.99, and 0.94 seconds per detection at input resolutions of 1802, 960, and 480 pixels, respectively. Each box is processed at a fixed internal resolution, so reducing the input resolution does not consistently reduce processing time. Generating descriptions can dominate total runtime in videos with many objects. These component timings do not measure end-to-end throughput or memory compression.

# C VQ3D LOCALIZATION AND DISCOVERY ANALYSIS

## C.1 PROTOCOL AND INTERPRETATION

We evaluate all 164 queries from the 44 Ego4D clips in the VQ3D validation split. Each target has two 3D annotations, $\mathbf{g}_1$ and $\mathbf{g}_2$, in the scan coordinate frame. We retrieve an object instance by its name and return its stored world position. We do not compare this with an otherwise identical system that retrieves objects from query crops.

We measure error relative to the midpoint of the two annotations:

$$L_2 = \left\| \hat{\mathbf{x}} - \frac{\mathbf{g}_1 + \mathbf{g}_2}{2} \right\|.$$

Conditional success uses the benchmark criterion

$$L_2 < 6\left(\|\mathbf{g}_1 - \mathbf{g}_2\| + 1\,\mathrm{m}\right).$$

This threshold is permissive. Combined with excluding unanswered queries, it produces near-ceiling conditional success even when localization errors are several metres. We therefore report $L_2$ rather than a leaderboard comparison, which requires success over all queries and the official $L_2$ aggregation from the benchmark's evaluation code. Appendix C.5 lists published results and the remaining protocol differences.

We score a query only when retrieval returns a candidate. Raw SAM 3 memory returns candidates for 141 of 164 queries, and trajectory memory for 151. Each variant is therefore scored on its own returned queries, rather than a shared set. A 99% conditional success rate or a median error below one metre does not imply successful localization on 99% of all queries. These results do not establish full-query success or an end-to-end improvement on the official VQ3D task.

## C.2 IMPLEMENTATION SETTINGS

We sample every twentieth frame from a 5 fps grid, giving one frame every four seconds. YOLO-World uses v8x-worldv2 at 640 pixels, with a confidence threshold of 0.25 and at most 12 detections per frame. We lift at most 25 detections per tag and 400 per scene into 3D. Trajectory association uses a 1.0 m radius, and consolidation uses a mask-IoU threshold of 0.3.

The localization tables use GPT-5.4 to select from at most four candidates. We evaluate Qwen only for tagging and judging coverage, so these results do not demonstrate a localization pipeline built entirely from open models. We also evaluate FastVGGT (Shen et al., 2025) as an alternative source of camera poses.

## C.3 LOCALIZATION ACROSS MEMORY VARIANTS

Raw memory groups observations by tag and returns the densest 3D cluster for the selected tag. Trajectory memory associates observations in time order using each instance's latest position, rather than a global centroid. This allows it to retain multiple instances with the same tag. Consolidation additionally groups synonymous tags and propagates SAM 3 masks to reconnect tracks, using a co-occurrence check to block incompatible merges. In this analysis, EgoLoc provides poses for approximately 98% of sampled frames, compared with approximately 100% for FastVGGT.

With EgoLoc poses, consolidation reduces the median by 0.029 m for SAM 3 and 0.026 m for YOLO-World. This task predicts a single position and does not score identity continuity, movement histories, or later scene-specific question answering. A small localization difference cannot establish that tracking is unnecessary for those tasks.

## C.4 TRIANGULATED AGAINST MONOCULAR POSITIONS

The triangulated variant intersects the viewing rays of an object's observations (Equation 1) with the HD-EPIC guards: at least three rays, a camera spread of 0.10 m, conditioning of 0.25, the point in front of every camera, and at most 2 m from the monocular estimate; otherwise the monocular position is kept. We compute viewing rays using the dataset's fisheye camera model. For each query,

Table 17: VQ3D localization on returned candidates. Error is in metres. Fractions and conditional success (CS) are percentages. Different rows may grade different queries. "Consolidated" adds mask tracking to trajectory memory.

| Detector | Memory | Poses | Median | Mean | $< 1$ m | $< 2$ m | CS |
|---|---|---|---|---|---|---|---|
| SAM 3 | Raw | EgoLoc | 0.877 | 1.399 | 53 | 77 | 99 |
| | Raw | FastVGGT | 1.064 | 1.504 | 45 | 76 | 99 |
| | Trajectory | EgoLoc | 0.984 | 1.388 | 52 | 78 | 99 |
| | Trajectory | FastVGGT | 1.057 | 1.449 | 45 | 77 | 99 |
| | Consolidated | EgoLoc | 0.955 | 1.360 | 53 | 78 | 99 |
| | Consolidated | FastVGGT | 1.057 | 1.421 | 45 | 76 | 99 |
| YOLO-World | Raw | EgoLoc | 1.090 | 1.526 | 45 | 71 | 100 |
| | Raw | FastVGGT | 1.268 | 1.627 | 38 | 66 | 100 |
| | Trajectory | EgoLoc | 1.089 | 1.475 | 48 | 73 | 99 |
| | Trajectory | FastVGGT | 1.177 | 1.587 | 39 | 69 | 99 |
| | Consolidated | EgoLoc | 1.063 | 1.441 | 48 | 74 | 99 |
| | Consolidated | FastVGGT | 1.141 | 1.509 | 42 | 71 | 99 |

GPT-5.4 selects one candidate, which we score using both its original and refined positions. This holds object selection fixed and avoids variation from repeated answering. The two ground-truth annotations differ by a median of 4.8 cm.

Table 18 reports results for queries that pass the refinement checks: about half with SAM 3 and a quarter with YOLO-World. Most remaining queries have fewer than three observations at this sampling rate. With EgoLoc poses, triangulation reduces error on eligible queries by a median of 0.09–0.12 m. All three SAM 3 results remain significant after Holm correction across twelve tests. With FastVGGT poses, the gain falls to about 0.03 m, showing that refinement depends on pose accuracy.

Across all graded queries, mean error decreases in every configuration, but median error does not improve. Some predictions with errors just below 1 m become slightly worse and move above that value, affecting the median. The share of predictions within 0.5 m nevertheless increases. The monocular medians here differ from those of Table 4 by up to 0.04 m because the matcher was re-run.

Table 18: VQ3D, triangulated against monocular positions, paired on the same retrieved object. $n$: queries whose object passes the triangulation guards; $\Delta$: median per-query change in $L_2$ (m; negative is closer) with bootstrap 95% interval; $p$: Wilcoxon signed-rank. The last two columns cover all graded queries, where objects that fail the guards keep the monocular position.

| Detector / level | Poses | $n$ | $\Delta$ median [95% CI] | $p$ | $<0.5$ m (%) | Mean $L_2$, all (m) |
|---|---|---|---|---|---|---|
| SAM 3/ raw | EgoLoc | 71 | $-0.09$ $[-0.18, -0.01]$ | 0.002 | $24 \to 31$ | $1.38 \to 1.34$ |
| SAM 3/ raw | FastVGGT | 67 | $-0.03$ $[-0.12, +0.01]$ | 0.029 | $22 \to 28$ | $1.53 \to 1.49$ |
| SAM 3/ trajectory | EgoLoc | 76 | $-0.11$ $[-0.18, -0.03]$ | 0.004 | $21 \to 32$ | $1.46 \to 1.42$ |
| SAM 3/ trajectory | FastVGGT | 67 | $-0.03$ $[-0.10, +0.02]$ | 0.064 | $24 \to 30$ | $1.54 \to 1.51$ |
| SAM 3/ tracker | EgoLoc | 70 | $-0.12$ $[-0.19, -0.03]$ | 0.003 | $27 \to 37$ | $1.36 \to 1.32$ |
| SAM 3/ tracker | FastVGGT | 69 | $-0.03$ $[-0.10, +0.02]$ | 0.112 | $25 \to 32$ | $1.43 \to 1.40$ |
| YOLO-World/ raw | EgoLoc | 38 | $-0.04$ $[-0.09, +0.02]$ | 0.199 | $24 \to 37$ | $1.53 \to 1.52$ |
| YOLO-World/ raw | FastVGGT | 36 | $-0.04$ $[-0.09, +0.04]$ | 0.315 | $19 \to 33$ | $1.62 \to 1.61$ |
| YOLO-World/ trajectory | EgoLoc | 31 | $-0.05$ $[-0.29, -0.00]$ | 0.058 | $16 \to 26$ | $1.50 \to 1.48$ |
| YOLO-World/ trajectory | FastVGGT | 33 | $-0.07$ $[-0.24, +0.00]$ | 0.018 | $18 \to 33$ | $1.61 \to 1.57$ |
| YOLO-World/ tracker | EgoLoc | 33 | $-0.05$ $[-0.14, +0.00]$ | 0.063 | $12 \to 21$ | $1.45 \to 1.43$ |
| YOLO-World/ tracker | FastVGGT | 31 | $-0.02$ $[-0.20, +0.01]$ | 0.102 | $16 \to 32$ | $1.49 \to 1.47$ |

### C.5 Published methods on the validation split

Five differences must be read with Table 19. (1) The benchmark identifies the target through a visual crop from the VQ2D response track. We instead give the object's text title to a language model that searches a pre-built memory, so the query input differs. (2) Recent methods report success rates of 96–99%, and our evaluation uses a fixed tolerance rather than one based on box size. Localization error ($L_2$) is therefore more informative than success rate, although the different query inputs still

Table 19: VQ3D validation split (164 queries, 44 clips), published methods and ours. Benchmark metrics as in Mai et al. (2023): Succ over all queries, Succ* over queries with a prediction, $L_2$ in metres, angular error in radians, QwP = queries with poses. Published rows are as reported. Our rows answer a text query from the memory, not the benchmark's image crop. n/a: angle not reported (published row) or not computed (our rows).

| Method | Succ (%) | Succ* (%) | $L_2$ (m) | Angle | QwP (%) |
|---|---|---|---|---|---|
| Ego4D baseline (Grauman et al., 2022) | 1.22 | 30.77 | 5.98 | 1.60 | 1.83 |
| Ego4D*, EgoLoc poses (Mai et al., 2023) | 73.78 | 91.45 | 2.05 | 0.82 | 80.49 |
| EgoLoc (Mai et al., 2023) | 80.49 | 98.14 | 1.45 | 0.61 | 82.32 |
| EgoLoc-v1 (Mai et al., 2024) | 81.13 | 98.10 | 1.45 | 0.55 | 84.73 |
| EAGLE (Cao et al., 2025) | 84.77 | 98.54 | **1.18** | **0.42** | 85.68 |
| Embodied VideoAgent (Fan et al., 2025) | 85.37 | 92.72 | 1.86 | n/a | 92.07 |
| Ours, SAM 3/ trajectory / EgoLoc | 92.7‡ | 99.3‡ | 1.41 (median 0.99) | n/a | 93.3§ |
| Ours, SAM 3/ consolidated / EgoLoc | 89.6‡ | 99.3‡ | 1.37 (median 1.00) | n/a | 90.2§ |
| Ours, YOLO-World/ trajectory / EgoLoc | 89.0‡ | 98.6‡ | 1.46 (median 1.06) | n/a | 90.2§ |
| Ours, SAM 3/ trajectory / FastVGGT | 93.9‡ | 99.4‡ | 1.49 (median 1.07) | n/a | 94.5§ |

‡Success is distance to the annotated midpoint below $6(\|\mathbf{g}_1 - \mathbf{g}_2\| + 1\,\mathrm{m})$. The official evaluation sets the success tolerance using the annotated box size, so our scores cannot be compared directly with published Succ values. Our tolerance is permissive enough that always returning the clip's mean camera position achieves 100% success (Table 20). §Not QwP: the fraction of queries for which the memory returned a located object (graded / 164).

limit direct comparison. (3) EgoLoc states its $L_2$ as an RMSE; the RMSE of our YOLO-World row is 1.80 m (mean 1.44, median 1.06), so our error is comparable if the published values are means and worse if they are RMSEs. (4) A small number of large errors raises our mean. With YOLO-World, 10% of queries have errors above 3 m and account for 26% of the total error. Excluding these queries lowers the mean to 1.19 m. Some misses may involve selecting the wrong instance of the same category, which a query crop could help distinguish. Errors also vary across clips: clip identity explains 42% of the $L_2$ variance. This is consistent with differences in scale accuracy, but does not establish scale error as the cause.

(5) EAGLE combines depth estimates across views using confidence weights and multi-view consistency, then aligns them to the scan with a similarity transform that includes scale. Our method instead combines per-frame monocular position estimates using an unweighted median. Angular error was not computed. Switching from EgoLoc to FastVGGT poses moves the YOLO-World mean from 1.44 to 1.51 m, with per-query errors correlated at 0.92, so poses are not the difference. We never submitted to the test server, which requires the image crop as input.

### C.6 Memories built by prior methods on the same clips

We build every memory on the same 44 clips and query it with the same text: "Where is the ⟨title⟩?" All methods use the same GPT-5.4 matcher and scoring procedure, and return positions in the scan coordinate frame. We use released code with default settings, changing only the adapters needed for this dataset.

ReMEmbR (Anwar et al., 2024) uses VILA (Llama-3-VILA1.5-8B) to caption three-second windows and stores the corresponding EgoLoc camera centres. It returns the camera position associated with the retrieved memory, as its original implementation does. This localizes the wearer rather than the object.

DirectMe (Wang et al., 2026a) builds its graph at 1 fps using the Objects365 vocabulary, YOLO-World, SAM 2, and Depth Anything 3. We align the graph to the scan using a similarity transform fitted to each clip's camera trajectory, with a camera RMS error of 0.58 m. The answerer receives no frames at query time.

OSNOM (Plizzari et al., 2024) uses our SAM 3 detections converted to VISOR-format masks, together with FastVGGT poses and depth. The estimated depth replaces its usual depth rendered from a COLMAP mesh. It returns the selected track's 3D location. AMEGO (Goletto et al., 2024) stores

hand–object interaction tracklets without object names or 3D positions, so it cannot answer these queries. The no-memory reference always returns the clip's mean camera position.

We re-evaluated our variants through this shared protocol on 2026-09-21. Their scores agree with the earlier results in Table 17 within answering variation. Table 21 compares methods on the 138 queries answered by every method capable of returning a position.

Table 20: All memories through one harness (164 validation queries). Succ and Succ* as in Table 19. The no-memory reference passes every success check, so $L_2$ and the sub-metre share are the informative columns. N.C.: cannot be computed, since AMEGO stores neither object names nor 3D positions.

| Memory | Graded | Succ | Succ* | Median | Mean | $< 1$ m | $< 2$ m |
|---|---|---|---|---|---|---|---|
| AMEGO | N.C. | N.C. | N.C. | N.C. | N.C. | N.C. | N.C. |
| ReMEmbR (wearer position) | 164 | 100.0 | 100.0 | 1.49 | 1.77 | 23 | 69 |
| DirectMe (graph only) | 164 | 99.4 | 99.4 | 1.91 | 1.98 | 18 | 52 |
| OSNOM (adapted) | 146 | 89.0 | 100.0 | 1.33 | 1.65 | 35 | 71 |
| *No memory: clip centroid* | 164 | 100.0 | 100.0 | 1.69 | 1.82 | 10 | 63 |
| SAM 3/ raw / EgoLoc | 143 | 86.6 | 99.3 | 0.95 | 1.41 | 52 | 77 |
| SAM 3/ raw / FastVGGT | 149 | 90.2 | 99.3 | 1.09 | 1.53 | 44 | 75 |
| SAM 3/ trajectory / EgoLoc | 153 | 92.7 | 99.3 | 0.99 | 1.41 | 51 | 76 |
| SAM 3/ trajectory / FastVGGT | 155 | 93.9 | 99.4 | 1.07 | 1.49 | 43 | 75 |
| SAM 3/ consolidated / EgoLoc | 148 | 89.6 | 99.3 | 1.00 | 1.37 | 50 | 78 |
| SAM 3/ consolidated / FastVGGT | 148 | 89.6 | 99.3 | 1.10 | 1.44 | 42 | 76 |
| YOLO-World/ raw / EgoLoc | 146 | 89.0 | 100.0 | 1.09 | 1.53 | 47 | 70 |
| YOLO-World/ raw / FastVGGT | 151 | 92.1 | 100.0 | 1.21 | 1.61 | 41 | 66 |
| YOLO-World/ trajectory / EgoLoc | 148 | 89.0 | 98.6 | 1.06 | 1.46 | 49 | 74 |
| YOLO-World/ trajectory / FastVGGT | 157 | 95.1 | 99.4 | 1.17 | 1.57 | 42 | 69 |
| YOLO-World/ consolidated / EgoLoc | 151 | 91.5 | 99.3 | 1.07 | 1.47 | 48 | 74 |
| YOLO-World/ consolidated / FastVGGT | 151 | 91.5 | 99.3 | 1.14 | 1.53 | 42 | 70 |

Table 21: The 138 queries answered by every method, which guards against methods answering different subsets.

| Memory | Median $L_2$ | Mean $L_2$ | $< 1$ m (%) | $< 2$ m (%) |
|---|---|---|---|---|
| LEDGER (SAM 3, trajectory, EgoLoc) | **1.00** | **1.44** | **50** | **76** |
| LEDGER (SAM 3, trajectory, FastVGGT) | 1.05 | 1.49 | 44 | 74 |
| OSNOM (adapted) | 1.33 | 1.65 | 34 | 72 |
| ReMEmbR | 1.48 | 1.74 | 22 | 71 |
| DirectMe (graph only) | 1.88 | 1.93 | 17 | 54 |
| Clip centroid (no memory) | 1.69 | 1.81 | 9 | 64 |

### C.7 DETECTOR AND VOCABULARY ANALYSIS

Table 22: Detector comparison with the same vocabulary, sampling, lifting, and pose settings. Localization uses trajectory memory and EgoLoc poses. Box IoU and 3D distance measure agreement with SAM 3 where both produce detections. SAM 3 is not ground truth. Timing covers detection only.

| Detector | s/frame | Detections/clip | Box IoU | 3D distance (m) | Median $L_2$ (m) |
|---|---|---|---|---|---|
| SAM 3 | 4.4–11.1 | 1,281 | – | – | 0.984 |
| YOLO-World | 0.051 | 424 | 0.928 | 0.439 | 1.089 |
| YOLOE-26 | 0.047 | 488 | 0.950 | 0.433 | 1.318 |
| YOLOE-11 | 0.049 | 471 | 0.961 | 0.415 | 1.404 |

SAM 3's reported time varies with the vocabulary (73–131 tags), whereas the YOLO variants batch prompts. These measurements do not include 3D lifting, consolidation, or descriptions, and are not

an end-to-end system speedup. Table 22 also shows that more detections or greater box agreement with SAM 3 need not correspond to lower downstream localization error.

Table 23: Reported semantic vocabulary coverage (%) of the 164 query objects. These are automatic label-coverage estimates, not measured object-instance recall. Vocabulary sizes are means per clip when available.

| Tag source | GPT judge | Qwen judge | Tags/clip |
|---|---|---|---|
| RAM++ | 53 | 39 | 53 |
| Gemma-3-4B | 76 | 64 | 297 |
| Qwen3.5-9B ∪ RAM++ | 84 | 68 | – |
| Canonical union | 86 | 70 | 112 |
| GPT vision, full | 87 | 76 | 177 |
| Gemma ∪ Qwen ∪ RAM++ | 87 | 75 | – |
| RAM++ ∪ GPT vision, full | 89 | 80 | 258 |

Coverage estimates vary substantially with the model used to judge whether the query object appears in the tag set. For Qwen tags, both judges accept 108 queries and reject 27. GPT alone accepts another 27, while Qwen alone accepts two. Of these disagreements, 19 in the first group and both in the second contain the exact query word among the tags. A semantic judge can therefore reject even an exact word match.

A matching tag only indicates vocabulary coverage. It does not show that the object was detected, that the correct instance was identified, or that its position was estimated correctly.

# D UCS-BENCH DETAILS

## D.1 BENCHMARK, CONSTRUCTION, AND ANSWERING PROTOCOL

UCS-Bench contains 8,114 timestamped questions across 532 videos from EgoLife, TeleEgo, HourVideo, EPIC-Kitchens, and EgoBlind. It provides no camera poses, intrinsics, or query boxes (Wang et al., 2026a). Most questions have five options; Proximity & Reachability questions have two, giving an overall chance accuracy of 25.7%. Our main evaluation uses 189 EgoLife recordings with 2,569 questions and 22 TeleEgo recordings with 203 questions. Tables 3 and 25 report the 2,771 questions shared across all conditions. We discuss HourVideo separately below.

Every comparison in the paper is paired on the questions that every arm in it answered, so the count changes with the arms compared; Table 24 lists the sets and where each one is reported.

Table 24: Which UCS-Bench questions each table uses. The benchmark holds 8,114 questions across 532 videos; we evaluate the EgoLife and TeleEgo recordings, and the stitched experiment (Appendix E.1) re-asks its questions on single recordings and again on the streams.

| Question set | $n$ | Reported in |
|---|---|---|
| *Single recordings, main evaluation* | | |
| EgoLife (189) + TeleEgo (22), shared by all conditions | 2,771 | Tables 3, 25 |
| the same recordings, arms of the cost comparison | 2,772 | Table 42 |
| with HourVideo (21 of 167 memories built) | 3,130 | Table 26 |
| *Stitched experiment, 206–212 recordings* | | |
| Single-recording arm, all seven conditions | 3,264 | Table 37; Figure 5b |
| after excluding API errors, paired single vs. stitched | 3,244 | Tables 28, 29 |
| answered by every answering model | 3,300 | Table 34 |
| answered by both records under both readers | 2,777 | Table 33 |
| answered by every leave-one-out arm | 2,246 | Table 32 |
| about an earlier scene, re-asked at the end | 763 | Table 35 |

**Memory construction.** We sample one frame every four seconds and use the same pipeline as on HD-EPIC: Qwen3.5-9B for tagging, YOLO-World for detection, WildDet3D for 3D lifting, SAM 3 for consolidation, followed by persistence segmentation and segment descriptions.

We use Depth Anything 3 poses when they pass three consistency checks: a chunk-alignment residual of at most 0.5 m, chunk scales within $[0.8, 1.25]$, and camera steps of at most 6 m. Otherwise, we use FastVGGT. Circular-fisheye recordings go directly to FastVGGT. We attempt metric scale recovery using pairs of observations of the same object. This fails for most recordings and all TeleEgo recordings, so their distances remain in scene units.

For each question, we filter the memory at its timestamp. We remove later observations, segments, and descriptions, truncate the current segment and wearer path, and recompute positions. We also withhold object summaries written using all segments. The adapter ranks entries by

$$3n_{\text{name}} + n_{\text{desc}} + 2[\text{near question time}] + \frac{1}{1 + \Delta t/60},$$

where $n_{\text{name}}$ and $n_{\text{desc}}$ count question-word matches in names and descriptions, and $\Delta t$ is the time gap to the nearest observation in seconds. The answerer receives the top four entries and the wearer's path.

**Answering conditions.** Every condition uses GPT-5.4 with one call per question. *Blind* receives only the question and options. *Memory* also receives the filtered memory as text. *Frames* receives 48 frames at 448 pixels, sampled uniformly from the recording's start to the question time. The benchmark uses 64 frames; we use 48 because this is the API limit.

*Frames + memory* adds memory text on spatial questions identified by a keyword rule, covering about 55% of questions. Adding memory to every question gives no gain over frames alone ($-0.8$ points, n.s.), and a split-roles prompt does not improve this result. When the API's image filter rejects frames, we fall back to the text condition. This affects 10–18% of EgoLife questions; restricting evaluation to questions whose images are never rejected gives the same differences.

We run DirectMe (Wang et al., 2026a) with its released graph builder and retrieval, using the Objects365 vocabulary, YOLO-World, SAM 2, and Depth Anything 3 at 1 fps. Its memory-only condition receives the graph as text. We test frames plus graph both with the graph always included, as in its original configuration, and with the same spatial-question routing used for our memory. DirectMe's own prompt without the graph scores 31.0%. Its single-threaded builder scales superlinearly with graph size, taking 7–40 hours for a 40-minute stream. We therefore use a batched wrapper, verified to produce identical output.

Table 25: UCS-Bench accuracy (%) by subcategory on the 2,771 questions of Table 3, with its abbreviations.

| | | | **Text only** | | | | **Frames +** | | |
|---|---|---|---|---|---|---|---|---|---|
| Subcategory | $n$ | Blind | Ours | DM graph | DM blind | Frames | Ours | DM, own | DM, routed |
| Quantity Change Tracking | 624 | 34.5 | 33.3 | 17.5 | 21.3 | 37.8 | 38.8 | 30.9 | 37.7 |
| My Trajectory | 465 | 35.9 | 41.3 | 38.3 | 43.2 | 45.8 | 48.8 | 41.5 | 44.1 |
| Relative Position & Orientation | 392 | 24.0 | 32.4 | 29.1 | 22.2 | 37.8 | 40.6 | 35.7 | 36.2 |
| Distance Comparison | 313 | 41.9 | 44.7 | 38.3 | 36.1 | 48.9 | 48.2 | 43.1 | 43.8 |
| Ego-centric Position | 309 | 23.6 | 34.6 | 22.7 | 23.6 | 42.1 | 48.2 | 39.5 | 38.8 |
| Object Category Recall | 260 | 32.7 | 42.3 | 30.8 | 30.0 | 55.0 | 53.5 | 46.9 | 52.3 |
| Reachability Judgment | 225 | 50.2 | 55.6 | 42.2 | 44.0 | 55.6 | 55.6 | 52.9 | 55.1 |
| Object Trajectory | 183 | 32.2 | 31.7 | 35.0 | 41.0 | 44.3 | 49.7 | 42.6 | 49.2 |
| All | 2771 | 33.8 | 38.5 | 30.0 | 31.0 | 44.4 | 46.3 | 39.8 | 42.9 |

**Paired tests.** We compare conditions on the same questions using exact McNemar tests. Our memory outperforms DirectMe's graph by 8.6 points ($p < 10^{-12}$) and the blind control by 4.7 points ($p < 10^{-4}$). DirectMe's graph scores 1.0 point below its own blind prompt ($p = 0.32$).

With frames, our memory outperforms DirectMe's graph by 3.4 points under the same routing ($p = 0.0005$) and improves on frames alone by 1.9 points ($p = 0.047$). Adding DirectMe's graph lowers accuracy relative to frames alone by 1.4 points with our routing ($p = 0.13$) and 4.6 points in its original configuration ($p < 10^{-5}$).

DirectMe uses a deterministic counting rule on 297 of the 2,771 questions, answering only 4.7% correctly. Excluding these questions, our memory still outperforms its graph by 6.1 points without frames.

Repeated frame-based runs on identical inputs change 16–20% of selected answers. On an earlier set of 1,581 questions, averaging two runs gives a frames-plus-memory gain over frames alone of 0.9 points, with an interval of $[-1.1, +3.0]$. These checks suggest that single-run differences below roughly 1.5 points may reflect answering variation.

**The published setting.** The UCS-Bench paper reports DirectMe with Qwen3-VL-8B, 64 frames, and graph plus keyframes on the full 8,114-question benchmark at 50.5% overall (54.1 / 47.6 / 58.0 / 46.4 by category) against 44.0% for the same model with frames only (44.6 / 40.1 / 50.9 / 42.9); it reports no graph-only number. Those numbers use a different answerer and question set and are not comparable with Table 3. The paper's setting with Qwen3-VL was not rerun.

Table 26: UCS-Bench accuracy (%) by source dataset. HourVideo covers only the 21 of 167 recordings whose memory could be built.

| Dataset | $n$ | Blind | Memory | 48 frames | Frames + memory |
|---|---|---|---|---|---|
| EgoLife (189 recordings) | 2,569 | 33.6 | 38.4 | 44.1 | 45.8 |
| TeleEgo (22 recordings) | 203 | 37.4 | 39.4 | 47.3 | 52.2 |
| HourVideo (21 of 167 recordings) | 358 | 39.7 | 41.1 | 53.1 | 49.4 |
| All three | 3,130 | 34.5 | 38.8 | 45.4 | 46.6 |

**By dataset and recording length.** Paired against blind, the memory gains $+4.9$ on EgoLife ($p < 10^{-4}$), $+2.0$ on TeleEgo (n.s.), and $+1.4$ on HourVideo (n.s.); pooled $+4.3$ $[+2.3, +6.3]$ (Table 26). On TeleEgo the questioned noun is absent from the memory as a track for 40 to 60% of questions and every pose set is unscaled. On HourVideo (20 to 96 minutes, median 30) only 21 of 167 memories were built: the tracking stage scales worse than linearly with the number of instances, 50 to 60-minute recordings hold about 4,000, and the measured cost reached 74 to 3,500 hours per recording, so the build was stopped. On the 21 built, adding the memory to the frames costs 3.6 points ($p = 0.20$), with $-16.5$ on Relative Position ($p = 0.004$), the opposite of EgoLife, where the memory adds 5 to 6 points on the two egocentric-geometry subcategories.

### D.2 Counting

Table 27: Quantity Change Tracking on EgoLife and TeleEgo (625 questions). Question types by keyword: objects present (336), the wearer's actions (168), people (119).

| Evidence | All | Objects | Actions | People |
|---|---|---|---|---|
| Blind | 34.6 | 37.2 | 28.0 | 36.1 |
| Memory, text | 33.3 | 33.6 | 29.8 | 37.0 |
| Memory, text + counting aid (tracks, most visible together) | 26.9 | 25.0 | 25.0 | 34.5 |
| 48 uniform frames at 448 px | 37.9 | 33.9 | 41.1 | 45.4 |
| 48 frames chosen from memory keyframes | 37.4 | 33.6 | 36.9 | 49.6 |
| 24 frames at 896 px, half from the last 24 s, + 6 keyframes | 39.5 | 38.4 | 42.3 | 37.8 |

No use of the memory helps counting (Table 27), and showing its counts to the answerer costs 6.4 points ($p < 0.001$). We evaluate counts on 257 questions with numeric answers whose target noun appears in memory. Counting tracks gives the correct answer on 11% of questions and is within one on 22%. The median estimate is seven objects, compared with a true count of three. Using the largest number of tracks visible in a single frame improves these rates to 16% and 45%, respectively.

The memory both splits one object into several tracks and misses other objects. On recordings without metric scale, the consolidation step's fixed 3D-distance threshold blocks 98% of proposed merges. Removing this check reduces the number of stored objects from 6,490 to 4,404 across 12 recordings, without lowering measured overall accuracy. Counting remains unresolved: the two configurations score 33.7% and 31.6% on 95 counting questions, both below the 35.8% blind score. Frames help with action and people counts, but object counts remain near the blind baseline. These results point to instance association and consistent geometric scale during construction as priorities for improving object counts.

### D.3 Calibration set and causal construction

We developed the adapter on a small calibration set of ten videos with 138 questions. These results are therefore not a held-out evaluation. The first version scored 36.2%, compared with 25.4% for the blind control, but included future evidence. On 57 questions, it retrieved a candidate first observed after the question time, and 43.6% of the displayed path waypoints were in the future. We exclude this gain from our claims.

The timestamp filtering described above removes explicit future evidence from the prompt. On the calibration set, the filtered memory scores 38.4% with GPT-5.4, compared with 31.9% blind: a gain of 6.5 points with an interval of $[-2.9, +15.9]$. Direction labels, relation episodes, and place context provide no measurable additional gain.

However, construction still uses the full recording. The vocabulary draws on frames throughout the video, and pose estimation, consolidation, and segmentation all run offline. Filtering their outputs does not show that earlier memory states are independent of later frames. To establish this, we would need to build every stage using only footage available by each query time and verify that appending footage does not change those earlier states. We therefore describe these runs as timestamp-filtered, rather than as a validated streaming evaluation.

Repeating GPT-5.4 on identical inputs changes the selected option on about a quarter of calibration questions. Our main claims therefore rely on paired comparisons across thousands of questions, rather than these calibration estimates.

### D.4 Implementation settings

Tagging uses at most 80 frames per recording, with a budget of 130 tokens per frame. Detection runs at 1280 pixels with a confidence threshold of 0.25 and at most 40 detections per frame. Lifting uses the HD-EPIC budgets. Consolidation uses a mask-IoU threshold of 0.3 and a 3D-distance limit of 1.0. Persistence requires three observations outside a radius of 0.3. Both distance thresholds are in scene units when metric scale is unavailable.

GPT-5.4 answers each question in one call, using about 1,500–1,800 input tokens with memory and 500 for the blind control. Depth Anything 3 estimates poses in 48-frame chunks with 12 overlapping frames. FastVGGT uses 96-frame chunks with 16 overlapping frames and joins them by Sim(3) alignment. If a chunk cannot be aligned, it starts from a new anchor instead of failing. This fallback was needed for scene-change streams, where some alignment errors at the boundary reached 3–21 scene units.

## E Multi-Scene UCS-Bench Streams

### E.1 Stream design

We construct 100 streams from 212 UCS-Bench recordings, using each recording once. Sixty same-scene streams combine two 12-minute EgoLife recordings of the same house on different days. Forty scene-change streams combine a 1–12-minute EPIC-Kitchens or EgoBlind clip with a 20-minute TeleEgo recording; twelve contain three parts. In these streams, the place, wearer, and camera change at the recording boundary.

We build one memory and one DirectMe graph over each complete stream. Tagging, pose estimation, association, and description generation therefore operate across recording boundaries. We ask every source question again, shifting its timestamp to its position in the stream. Frame-based conditions sample from the stream's start up to the question time.

The same questions were answered on their original recordings two to five days earlier using the same code. Between these runs, blind accuracy changes by 0.3 points and 27% of selected options differ, providing a check on answering variation. Of 3,264 questions paired across all seven conditions, 3,244 remain after excluding questions affected by API errors. We recorded 23 such errors and excluded them rather than scoring them as incorrect. Image-refusal rates are similar: 11.4% on individual recordings and 11.5% on stitched streams. Table 28 gives paired accuracy for every condition.

The protocol changes recording length as well as scene identity: scene-change streams place a short clip before a longer recording. Their effects therefore cannot be separated in this comparison. An earlier experiment on 303 streams merged memories built separately for each recording and showed little accuracy change ($-1.6$ and $+0.9$ points). That setup kept construction within each recording; constructing memory over the joined stream exposes failures that merging completed memories does not test.

Table 28: Paired accuracy (%) on individual recordings and 100 stitched streams (3,244 questions). Same scene joins different days in one house; scene change joins different places. $\Delta$: stitched minus single; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (McNemar). The last row emulates per-scene construction with detected boundaries (Appendix E.4).

| | **All** (3,244) | | | **Same scene** (1,959) | | | **Scene change** (1,285) | | |
|---|---|---|---|---|---|---|---|---|---|
| Evidence | single | stitched | $\Delta$ | single | stitched | $\Delta$ | single | stitched | $\Delta$ |
| None (blind) | 33.2 | 33.5 | $+0.3$ | 33.0 | 32.9 | $-0.2$ | 33.4 | 34.4 | $+1.0$ |
| 48 frames | 47.8 | 44.2 | $-3.6^{***}$ | 44.0 | 41.2 | $-2.7^{*}$ | 53.7 | 48.6 | $-5.1^{***}$ |
| DM graph | 30.1 | 29.6 | $-0.5$ | 29.7 | 29.7 | $-0.1$ | 30.7 | 29.6 | $-1.2$ |
| DM frames + graph (own) | 42.9 | 40.9 | $-2.0^{*}$ | 38.9 | 39.9 | $+1.0$ | 48.9 | 42.4 | $-6.5^{***}$ |
| LEDGER memory | 39.0 | 36.4 | $-2.5^{**}$ | 38.4 | 37.9 | $-0.5$ | 39.8 | 34.2 | $-5.7^{***}$ |
| LEDGER frames + memory | 48.5 | 45.3 | $-3.3^{***}$ | 45.4 | 44.5 | $-0.9$ | 53.2 | 46.4 | $-6.8^{***}$ |
| ReMEmbR memory | 41.4 | 40.9 | $-0.5$ | 41.3 | 39.7 | $-1.6$ | 41.5 | 42.7 | $+1.2$ |
| LEDGER memory + event lines | 41.0 | 39.2 | $-1.8^{*}$ | 39.8 | 39.4 | $-0.4$ | 42.7 | 38.9 | $-3.8^{**}$ |
| LEDGER memory, per-scene build | 38.9 | 37.4 | $-1.5^{*}$ | 38.3 | 37.6 | $-0.7$ | 39.8 | 37.0 | $-2.8$ |

Table 29: Stitched $-$ single (points) by the position of the question's recording in the stream and by stream type. Stars as in Table 28; cells whose recorded $p$ is rounded to 0.00 are marked $^{**}$.

| Subset | $n$ | Blind | Frames | Ours mem. | Ours fr.+mem. | DM graph | DM fr.+graph | DM fr.+graph, routed |
|---|---|---|---|---|---|---|---|---|
| All | 3244 | $+0.3$ | $-3.6^{***}$ | $-2.5^{**}$ | $-3.3^{***}$ | $-0.5$ | $-2.0^{*}$ | $-3.4^{***}$ |
| First recording | 1503 | $+0.7$ | $-0.2$ | $-0.9$ | $+0.7$ | $-1.7^{*}$ | $-0.9$ | $-0.3$ |
| After one earlier scene | 1558 | $-0.4$ | $-5.6^{**}$ | $-3.0^{*}$ | $-6.0^{**}$ | $-0.3$ | $-2.6$ | $-6.1^{**}$ |
| After two earlier scenes | 183 | $+3.3$ | $-15.3^{**}$ | $-11.5^{*}$ | $-12.0^{*}$ | $+7.7^{*}$ | $-6.0$ | $-6.0$ |
| Scene-change streams | 1285 | $+1.0$ | $-5.1^{***}$ | $-5.7^{***}$ | $-6.8^{***}$ | $-1.2$ | $-6.5^{***}$ | $-6.1^{***}$ |
| Same-scene streams | 1959 | $-0.2$ | $-2.7^{*}$ | $-0.5$ | $-0.9$ | $-0.1$ | $+1.0$ | $-1.6$ |

### E.2 WHERE THE LOSS APPEARS

Questions in a stream's first recording are unchanged for every arm (Table 29); the loss grows with the number of earlier scenes. The loss on scene-change streams varies by the question's source recording. EPIC-Kitchens questions lose 8.0 points with memory ($p < 10^{-4}$) and 8.5 with frames. Memory losses on TeleEgo and EgoBlind are smaller, at 2.5 and 1.4 points, respectively, and neither is significant.

Timestamp-filtered retrieval still returns objects first seen in an earlier recording for 29% of EPIC-Kitchens questions and 4% of TeleEgo questions. Accuracy falls from 42% to 24% on questions with such a candidate, compared with 41% to 37% on questions without one. On same-scene streams, these objects come from the same kitchen, and we observe no corresponding drop (38.9% to 39.9%). Across question categories, the memory's consistent decline is in Proximity & Reachability, from 51.0% to 45.1%.

On stitched streams, our memory outperforms DirectMe's graph by 6.8 points overall ($p < 10^{-9}$), or 4.0 points after excluding questions handled by DirectMe's counting rule. The advantage is 4.6 points on scene-change streams. With frames and the same memory-routing rule, our memory outperforms DirectMe's graph by 3.1 points overall ($p = 0.0009$) and 0.4 points on scene-change streams.

Compared with frames alone, adding our memory gains 1.1 points on stitched streams, which is not significant, and 0.7 points on individual recordings.

Table 30: Scene-change streams, memory-only arm, 1,285 paired questions: which change recovers the loss. "True cuts" use the known seam times; "detected cuts" come from the video.

| Memory over the stitched stream | Accuracy | vs. single-scene build ($p$) |
|---|---|---|
| Single-recording build (reference) | 39.8 | $+5.7$ ($< 0.001$) |
| Single-scene build over the stream | 34.2 | – |
| + retrieval restricted to the current scene, true cuts | 35.3 | $+1.1$ (0.37) |
| + retrieval restricted, cuts from the place-partition detector | 35.0 | $+1.0$ (0.34) |
| Per-scene construction, true cuts | 38.4 | $+4.3$ ($< 0.01$) |
| Per-scene construction, detected cuts | 37.0 | $+2.9$ (0.05) |

### E.3 Construction and retrieval

Using the true scene boundaries to restrict retrieved candidates recovers 1.1 accuracy points ($p = 0.37$; Table 30). Also trimming the displayed histories to the current scene increases the recovery to 3.0 points ($p = 0.011$) on the same questions. This supports cross-scene retrieval as one source of the loss.

Memory construction introduces two further problems. First, tag canonicalization assumes a single physical scene and drops labels that appear inconsistent with it. In one stitched stream, it retains 98 distinct labels, compared with 121 across the original recordings. For 96% of questions after a scene boundary, the retrieved candidates differ from those returned by the separate builds; their label multisets have a Jaccard similarity of 0.3.

Second, the Depth Anything 3 pose check can reject an entire stream because of an error at the boundary. One boundary residual is 0.61 m, exceeding the 0.5 m limit. Across 29 of 37 scene-change streams, the pipeline falls back to unscaled FastVGGT poses. The number of metric-distance mentions in the prompts falls from 2,305 to 856. Neither failure occurs in the same-scene streams.

Building memory separately for each recording brings accuracy closer to the single-recording results at every stream position: 37.0% versus 37.2% for questions in the first recording, 40.4% versus 42.9% after one earlier scene, and 36.6% versus 38.3% after two. Recovery is therefore substantial but incomplete.

### E.4 Scene-cut detection and per-scene construction

Cuts are detected by a no-revisit test on CLIP embeddings of the 1 fps frames: $r_{10}$ is the 10th percentile of the distance from the next five minutes of frames to their nearest earlier frame, divided by the median nearest-neighbour distance among earlier frames at least 30 s apart, and a cut is declared where $r_{10} \geq 1$. Across the 100 streams, the detector finds 89 of 112 true boundaries within 60 s, including 42 of 52 on scene-change streams. It also adds 2.3 extra cuts per stream, mostly at room changes within EgoLife recordings, and 1.1 per individual recording. These extra cuts have little measured effect on same-scene accuracy (37.6% versus 37.9%). Grayscale thumbnails and colour histograms did not reliably distinguish boundaries, while a place-label consistency filter removed genuine cuts.

We first emulate per-scene construction using existing per-recording memories. A detected segment contained within one recording uses that recording's memory restricted to the segment's time window. A segment spanning a missed boundary uses the memory built over the joined stream. Of 422 detected segments, 61 coincide with an entire recording.

Across all streams, this emulated memory scores 37.4%, compared with 36.4% when building one memory over each stream ($+1.0$ points, $p = 0.29$). It remains below the 38.9% score on individual recordings ($-1.5$ points, $p = 0.048$). For questions after two earlier scenes, the gain over the whole-stream build is 6.6 points ($p = 0.13$).

We also run the full pipeline separately on each detected segment, including tag canonicalization, pose checks, and scale recovery. We evaluate this rebuild on 39 of the 40 scene-change streams, covering 1,251 questions; one stream is excluded because a segment has no usable tags. Accuracy reaches 36.5%, compared with 33.8% for the whole-stream build ($+2.7$ points, $p = 0.08$) and 39.5%

for individual-recording builds. Gains are 4.4 points after one earlier scene and 7.1 after two. These results are consistent with the partial recovery estimated by emulation.

### E.5 EVENT LINES

For each sampled frame, Qwen3.5-9B writes an event line of about 18 words describing what the wearer is doing, the objects or people involved, and the location. We store these lines alongside the object records. At answer time, the reader receives the last six lines before the question and the ten that best match its words.

We developed this extension on 22 of the 100 stitched streams (807 questions). Unless stated otherwise, the main text reports results on the remaining 78 streams (2,461 questions). On these held-out streams, event lines improve accuracy by 2.0 points $[0.5, 3.6]$ on individual recordings ($p = 0.023$) and 2.6 points $[0.9, 4.4]$ on stitched streams ($p = 0.004$).

We also tested longer, 60-word lines and nine retrieval or prompt variants. Only one variant improved calibration accuracy ($+4.2$ points), and that gain did not carry over to the held-out streams. All ablations below use one GPT-5.4 run per condition, evaluated on the questions answered by every condition in the table. We use exact McNemar tests; overall accuracy varies by about $\pm 1$ point across repeated runs.

The same event lines were built on the HD-EPIC frames (36,633 lines, one per sampled frame) and shown to the answerer as the 16 lines that best match the question, together with the names of the retrieved candidates, since HD-EPIC questions mark objects by box. Table 31 compares them with the same memory re-answered in the same session. The mean moves by $+0.7$ $[-1.3, +2.5]$, no prototype changes significantly, and the gap to AMEGO on fixture-interaction counting and stationary localization remains.

Table 31: Event lines on HD-EPIC (2,400 questions, GPT-5.4): accuracy (%) per prototype for the memory re-answered in the same session (control) and with event lines; $\Delta$ is paired, with 95% bootstrap intervals. AMEGO (adapted) is shown for reference.

| Prototype | Control | Event lines | $\Delta$ [95% CI] | AMEGO |
|---|---|---|---|---|
| Object location | 48.8 | 50.8 | $+2.0$ $[-1.6, +5.6]$ | 39.8 |
| Object contents retrieval | 30.5 | 35.0 | $+4.5$ $[-1.5, +10.5]$ | 22.0 |
| Fixture location | 56.2 | 58.0 | $+1.8$ $[-1.0, +4.6]$ | 12.8 |
| Fixture interaction counting | 41.3 | 40.7 | $-0.7$ $[-4.3, +3.0]$ | 47.0 |
| Movement itinerary | 40.4 | 42.2 | $+1.8$ $[-2.4, +5.8]$ | 26.8 |
| Movement counting | 42.5 | 43.0 | $+0.5$ $[-4.5, +5.5]$ | 34.5 |
| Stationary object localization | 36.0 | 32.0 | $-4.0$ $[-11.5, +3.5]$ | 55.5 |
| Mean (prototype-averaged) | 41.9 | 42.6 | $+0.7$ $[-1.3, +2.5]$ | 34.7 |

### E.6 LEAVE-ONE-OUT ABLATION

Table 32 removes one part of the memory with event lines at a time (Figure 7a). No removal moves the overall score by more than 1.3 points on the questions every arm answered, whereas the changes by category reach 13 points (egocentric position without 3D geometry) and differ in sign across categories. On the larger set of questions answered by both the reference and a given arm (3,103 to 3,171), removing the event lines costs 2.2 points ($p = 0.006$) and the descriptions 1.8 ($p = 0.02$); the other removals stay within noise.

### E.7 RECORD AGAINST READER

ReMEmbR (Anwar et al., 2024) captions every sampled frame with VILA, stores the camera position and time of each caption, and answers with an agent that issues embedding searches over the captions in several rounds, about four calls and 23k tokens per question; our reader is one keyword lookup and one call, about 3 to 5k tokens. Table 33 crosses the two records with the two readers (Figure 7b). To be indexed by ReMEmbR's embedder, our record is re-expressed as one text item

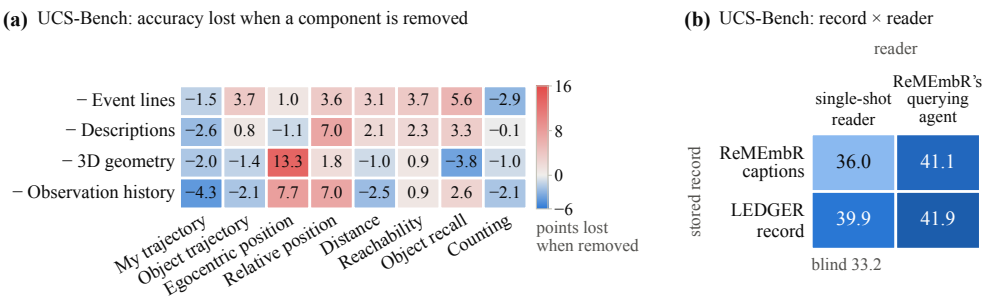


Figure 7: UCS-Bench stitched streams, GPT-5.4. (a) Accuracy lost on each question category when one component is removed from the memory with event lines (single runs, 2,246 questions answered by every arm; Table 32). (b) The same two records read by our single-shot reader and by ReMEmbR's querying agent (2,777 paired questions; Table 33).

Table 32: Leave-one-out on the stitched streams (2,246 questions answered by every arm). Category columns: my trajectory, object trajectory, egocentric position, relative position, distance comparison, reachability, object recall, counting.

| Memory | All | MyTraj | ObjTraj | EgoPos | RelPos | Dist | Reach | Recall | Count |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | 2,246 | 348 | 138 | 195 | 329 | 198 | 218 | 268 | 552 |
| Full (objects + event lines) | 39.0 | 35.6 | 39.9 | 37.4 | 40.7 | 47.0 | 57.8 | 42.5 | 28.3 |
| − event lines | 37.8 | 37.1 | 36.2 | 36.4 | 37.1 | 43.9 | 54.1 | 36.9 | 31.2 |
| − object descriptions | 37.6 | 38.2 | 39.1 | 38.5 | 33.7 | 44.9 | 55.5 | 39.2 | 28.4 |
| − 3D geometry | 38.6 | 37.6 | 41.3 | 24.1 | 38.9 | 48.0 | 56.9 | 46.3 | 29.3 |
| − observation history | 38.4 | 39.9 | 42.0 | 29.7 | 33.7 | 49.5 | 56.9 | 39.9 | 30.4 |
| − relations | 39.9 | 37.9 | 46.4 | 37.4 | 38.9 | 49.5 | 58.7 | 41.8 | 29.2 |
| − places | 39.2 | 37.4 | 42.0 | 37.9 | 38.3 | 47.0 | 52.8 | 41.8 | 31.2 |
| − objects (event lines only) | 38.1 | 40.2 | 42.8 | 24.6 | 35.3 | 47.5 | 57.8 | 40.3 | 29.7 |

per sampled moment (event line, place, and the objects in view with their descriptions and positions); adding each object's direction and distance from the wearer to the items did not help. We use ReMEmbR's captions read by its own agent as the reference. On paired questions, switching those captions to our reader lowers accuracy by 5.0 points ($p < 0.001$, exact McNemar test). Our records score 1.2 points below the reference with our reader ($p = 0.35$) and 0.8 points above it with ReMEmbR's agent ($p = 0.42$).

ReMEmbR's captions still outperform the blind control under our reader by 2.6 points ($p = 0.01$), showing that they contain useful evidence. However, their advantage under the native-system comparison does not persist when both memories use the same reader. Routing spatial questions to our reader and all remaining questions to ReMEmbR's agent yields 42.0% accuracy on 3,357 questions.

Table 33: Record against reader on the stitched streams (2,777 questions answered by every arm). R1 is our single-shot reader; R2 is ReMEmbR's retrieval agent. Blind on these streams is 33.2.

| Record → reader | All | MyTraj | ObjTraj | EgoPos | RelPos | Dist | Reach | Recall | Count |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | 2,777 | 423 | 156 | 270 | 409 | 288 | 257 | 307 | 667 |
| ReMEmbR captions → R1 | 36.0 | 38.5 | 37.8 | 29.6 | 32.3 | 41.7 | 55.3 | 32.9 | 30.6 |
| ReMEmbR captions → R2 (= ReMEmbR) | 41.1 | 45.9 | 50.0 | 28.1 | 37.9 | 42.7 | 48.6 | 44.3 | 37.9 |
| Our event lines only → R1 | 38.2 | 40.2 | 39.7 | 26.7 | 33.5 | 46.2 | 58.0 | 41.4 | 31.8 |
| Our objects + ReMEmbR captions → R1 | 38.5 | 39.5 | 36.5 | 35.6 | 38.9 | 45.8 | 55.6 | 39.7 | 29.1 |
| LEDGER (objects + event lines) → R1 | 39.9 | 37.1 | 40.4 | 36.7 | 40.6 | 45.8 | 58.8 | 45.6 | 30.0 |
| LEDGER → R2 | 41.9 | 49.2 | 47.4 | 30.4 | 35.9 | 45.1 | 54.5 | 44.6 | 36.7 |
| LEDGER + egocentric geometry → R2 | 41.3 | 47.0 | 48.1 | 25.6 | 34.2 | 44.1 | 55.6 | 45.6 | 38.2 |

### E.8 OPEN ANSWERING MODELS

Table 34 replaces GPT-5.4 by Qwen3.5-9B for our memory and its blind control, and runs DirectMe in its published configuration, Qwen3-VL-8B reading the graph with keyframes. On all

3,367 questions, with 58 unparsed Qwen3.5-9B answers scored as wrong, the memory's gain over the Qwen3.5-9B blind control is 3.4 points ($p < 0.001$).

Table 34: Answering models on the stitched streams (3,300 questions answered by every arm).

| Arm | All | MyTraj | ObjTraj | EgoPos | RelPos | Dist | Reach | Recall | Count |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | 3,300 | 511 | 191 | 305 | 480 | 318 | 308 | 374 | 813 |
| Qwen3.5-9B, blind | 32.3 | 39.5 | 46.6 | 22.3 | 28.5 | 38.1 | 49.0 | 34.2 | 20.9 |
| LEDGER + event lines, Qwen3.5-9B | 36.2 | 38.0 | 35.6 | 36.1 | 34.8 | 42.8 | 53.6 | 38.0 | 26.3 |
| DirectMe graph + keyframes, Qwen3-VL-8B | 28.0 | 33.3 | 33.5 | 24.9 | 27.1 | 36.5 | 46.8 | 26.2 | 15.5 |
| LEDGER + event lines, GPT-5.4 | 39.5 | 36.2 | 38.7 | 37.0 | 39.8 | 45.9 | 58.1 | 44.7 | 30.6 |
| ReMEmbR, GPT-5.4 | 40.8 | 45.0 | 50.3 | 27.5 | 38.5 | 42.5 | 49.4 | 43.6 | 37.0 |

### E.9 DELAYED QUESTIONS

The 763 questions about an earlier scene that do not depend on the present moment (object recall, counting, and object and wearer trajectories) are re-asked at the end of the stream, about their original moment, from the memory held at the end. Rewording a question as delayed moves the blind prior by $+5.5$ on the same questions, so Table 35 reports each arm's change net of blind: the frame reader loses most, since its 48 frames must cover the whole stream, and the memories lose 3 to 4 points. When the original moment is given exactly, rounded to 1, 5, or 10 minutes, or not at all, ReMEmbR is insensitive to the hint (40.9 to 42.7), $\text{LEDGER}_{\text{agent}}$ is within 1 to 2 points of it at every level (n.s.), and our single-shot reader is weaker (30.7 to 36.7) because its time bonus favours what was seen near the question, which for a delayed question is the end of the stream.

Table 35: Delayed questions (763): accuracy at the question's own moment and when re-asked at the end of the stream, with the change net of the blind control.

| Evidence | At its moment | Delayed | $\Delta$ | $\Delta$ net of blind |
|---|---|---|---|---|
| Blind | 32.4 | 37.9 | $+5.5$ | – |
| 48 frames | 44.0 | 39.6 | $-4.4$ | $-9.9$ |
| DirectMe graph | 25.9 | 28.3 | $+2.5$ | $-3.0$ |
| ReMEmbR | 40.6 | 42.7 | $+2.1$ | $-3.4$ |
| LEDGER, objects only | 34.1 | 32.8 | $-1.4$ | $-6.9$ |
| LEDGER + event lines | 35.1 | 36.6 | $+1.4$ | $-4.1$ |

## F ACCURACY BY RECORDING LENGTH

Figure 5 summarizes this appendix. Every arm is answered by GPT-5.4 and paired on the same questions, and a question is kept only if every arm in the table answered it; an invalid or missing answer counts as wrong. Accuracies here are over questions, so the HD-EPIC "All" row (44.6) differs from the prototype-averaged score of Table 2 (42.6). Lengths are the durations of the source videos. Intervals are Wilson intervals for accuracies and paired bootstrap intervals over videos (2,000 draws) for differences; the trend is a logistic regression of correctness on $\log_2$ of the recording length, so its slope is the change in log-odds per doubling. $\text{LEDGER}_{\text{agent}}$ is our memory read by ReMEmbR's querying agent (Appendix E.7); the "agent" columns below are this arm.

**HD-EPIC.** All 2,400 questions on 150 videos (Table 36). Every arm loses accuracy on longer recordings, and the blind control loses it fastest (Table 39), so the questions on long recordings are harder to guess. Our memory's gain over the blind control is $+9.9$ $[+2.5, +17.7]$, $+18.0$, $+16.2$, $+14.4$, and $+14.8$ $[+5.9, +23.8]$ points across the five bins, and its lead over ReMEmbR is 12 to 19 points in every bin. Length is partly a proxy for participant, since each participant records one kitchen, but every bin holds six to nine of the nine participants.

**UCS-Bench.** The 3,264 questions of the single-recording arm of the stitched experiment, on 206 recordings of up to 20 minutes (Table 37). No memory changes with length, and neither does the blind control; the 48-frame reader, the only arm that sees video, loses accuracy with every doubling

Table 36: HD-EPIC accuracy (%, over questions) by recording length (minutes).

| | | | LEDGER | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Length | Videos | Questions | single-shot | agent | ReMEmbR | AMEGO | DirectMe | OSNOM | Blind |
| <5 | 45 | 284 | **48.6** | 35.6 | 29.6 | 35.2 | 33.8 | 38.4 | 38.7 |
| 5–10 | 22 | 266 | **48.5** | 39.1 | 32.7 | 31.6 | 27.1 | 32.7 | 30.5 |
| 10–20 | 37 | 769 | **45.1** | 33.2 | 27.8 | 31.3 | 29.6 | 31.1 | 28.9 |
| 20–40 | 34 | 710 | **42.4** | 30.1 | 23.4 | 31.1 | 30.3 | 32.7 | 28.0 |
| 40–75 | 12 | 371 | **41.8** | 37.2 | 29.9 | 31.3 | 26.4 | 30.5 | 27.0 |
| All | 150 | 2,400 | **44.6** | 33.8 | 27.6 | 31.8 | 29.5 | 32.5 | 29.7 |

of length. The bins are not balanced across sources (the 11–13-minute bin is mostly EgoLife and holds 63% of the questions), so we repeat the fit within EPIC-Kitchens, the one source present in every bin (45 recordings, 801 questions): the frame reader still declines ($p = 0.05$) and no memory does (Table 39).

Table 37: UCS-Bench accuracy (%) by recording length (minutes), single recordings.

| | | | LEDGER | | | | | |
|---|---|---|---|---|---|---|---|---|
| Length | Videos | Questions | single-shot | agent | ReMEmbR | DirectMe | Blind | 48 frames |
| <2 | 16 | 236 | 36.0 | 41.5 | 37.7 | 31.4 | 33.1 | 61.4 |
| 2–6 | 23 | 370 | 38.1 | 43.5 | 42.4 | 28.6 | 32.4 | 50.0 |
| 6–11 | 18 | 321 | 45.8 | 49.5 | 46.1 | 29.9 | 34.9 | 55.1 |
| 11–13 | 123 | 2,050 | 38.6 | 40.4 | 40.8 | 29.3 | 32.6 | 44.5 |
| 13–20 | 26 | 287 | 37.3 | 44.3 | 41.5 | 36.2 | 35.2 | 48.1 |
| All | 206 | 3,264 | 39.0 | 42.1 | 41.4 | 30.1 | 33.1 | 47.7 |

**VQ3D.** The validation clips come in two lengths, about 5 and about 8 minutes, so this is not a length sweep (Table 38). The measure is the share of all 164 queries localized within 1 m, with an unanswered query counted as a miss.

Table 38: VQ3D: queries localized within 1 m (%) by clip length (minutes).

| | | | LEDGER | | | | | |
|---|---|---|---|---|---|---|---|---|
| Length | Clips | Queries | single-shot | agent | ReMEmbR | DirectMe | OSNOM | Clip centroid |
| ≈5 | 16 | 66 | 56.1 | 59.1 | 22.7 | 10.6 | 34.8 | 7.6 |
| ≈8 | 26 | 93 | 43.0 | 45.2 | 23.7 | 21.5 | 30.1 | 11.8 |
| 16 | 2 | 5 | 20.0 | 0.0 | 20.0 | 40.0 | 0.0 | 20.0 |
| All | 44 | 164 | 47.6 | 49.4 | 23.2 | 17.7 | 31.1 | 10.4 |

Table 39: Change in the log-odds of a correct answer per doubling of recording length, with its $p$-value in parentheses.

| Arm | HD-EPIC | UCS-Bench | UCS-Bench, EPIC-Kitchens only |
|---|---|---|---|
| LEDGER, single-shot reader | $-0.072$ (0.028) | $+0.013$ (0.69) | $+0.021$ (0.74) |
| $\text{LEDGER}_{\text{agent}}$ | $-0.054$ (0.12) | $-0.007$ (0.82) | $+0.022$ (0.72) |
| ReMEmbR | $-0.060$ (0.099) | $+0.016$ (0.63) | $+0.086$ (0.17) |
| DirectMe | $-0.072$ (0.044) | $+0.013$ (0.71) | $+0.010$ (0.89) |
| AMEGO (adapted) | $-0.076$ (0.029) | – | – |
| OSNOM (adapted) | $-0.084$ (0.016) | – | – |
| Blind | $-0.155$ ($<10^{-4}$) | $+0.008$ (0.82) | $-0.066$ (0.3) |
| ■48 frames | – | $-0.157$ ($<10^{-4}$) | $-0.122$ (0.05) |

# G FAILURE ANALYSIS AND QUALITATIVE EXAMPLE

## G.1 WHEN AN ANSWER IS WRONG, WHERE WAS IT LOST?

Table 40 follows the evidence through the pipeline. The largest measured losses happen before the answerer reads the record: on the calibration videos, giving GPT-5.4 the ground-truth trajectory of the queried object in place of the stored one adds 12 points, and naming the fixtures it rested on adds another 25, whereas a fixed rule improves on the answerer's reasoning by only 3 points on fixture location (Appendices B.7 and B.2). Each weak category of RQ1 (Section 4) maps to one row: stationary localization and fixture-interaction counting to what the record keeps about rest and hands, counting to association, and long recordings to construction.

Table 40: Where evidence is lost, by pipeline stage: what each stage loses, the questions it affects, and the measurement behind it.

| Stage | What is lost | Questions affected | Evidence |
|---|---|---|---|
| Sampling (one frame per 4 s) | objects seen only briefly; views for triangulation | all | 2 s sampling did not help the calibration set (App. B.6) |
| Tagging | objects outside the per-video vocabulary never enter memory | any about those objects | the best tag set misses about one VQ3D query object in ten (Table 23) |
| 3D lifting | depth along the viewing ray | metric localization | the error lies almost entirely along the ray (App. B.7) |
| Association | one object split into several tracks | counting | the track count is exact for 11% of UCS-Bench count questions (App. D.2) |
| Record content | continuous rest intervals; hand events | stationary localization, fixture-interaction counting | rest-segment starts are near chance; hand tracklets score higher (App. B.8, Table 10) |
| Construction | labels and metric scale at a scene cut; tractability with length | after a scene change; hour-long recordings | RQ5; 21 of 167 HourVideo memories built (App. D) |
| Retrieval | a wrong instance of the right category | the VQ3D error tail | 10% of queries are off by more than 3 m (App. C.5) |
| Reasoning | retrieved evidence left unused | fixture location | a fixed rule beats GPT-5.4 on the same evidence (App. B.2) |

## G.2 THREE MEMORIES ON ONE STREAM

Figure 8 shows DirectMe, ReMEmbR, and LEDGER answering from their own memories of one stitched stream of three EPIC-Kitchens recordings, and localizing a chair in a separate VQ3D clip. The questions were chosen to show how the memories differ, not how often; rates are in Tables 3 and 28. Our memory answers all four stream questions, DirectMe's graph answers one, and ReMEmbR none.

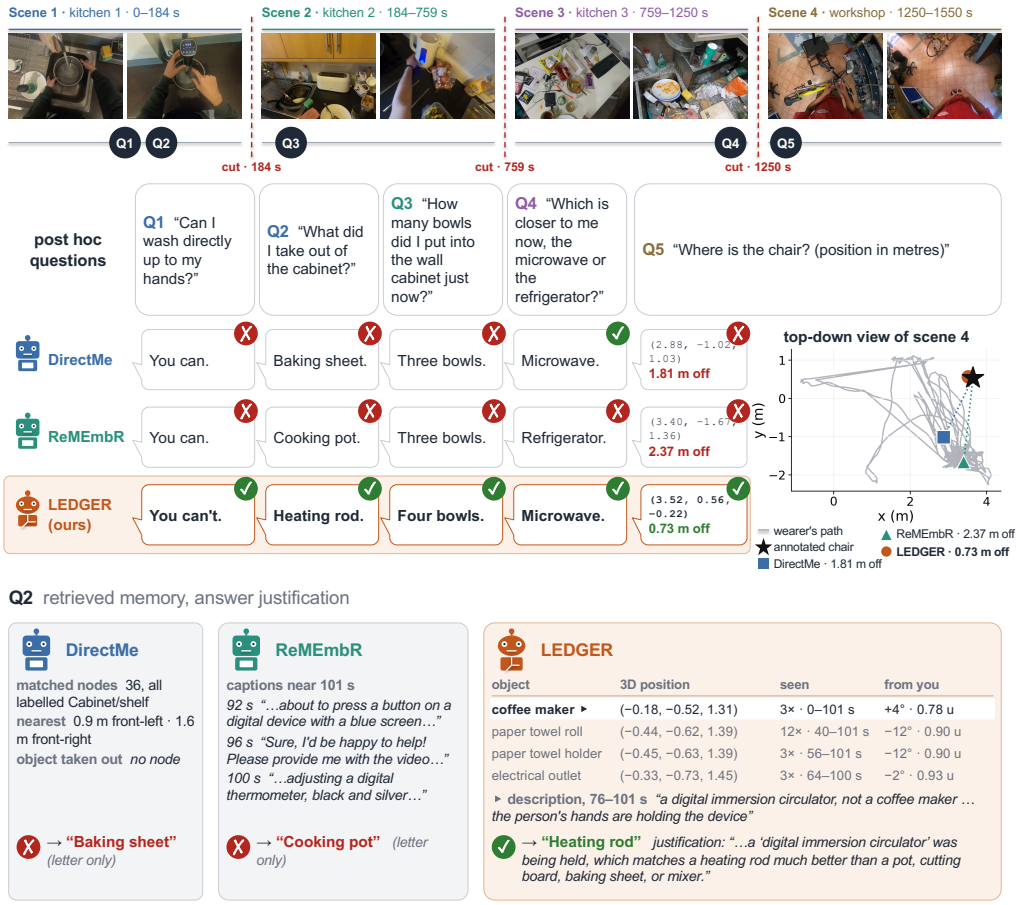


Figure 8: Three memories answer questions about a multi-scene egocentric stream from their stored records alone. Scenes 1–3 form one UCS-Bench stream that joins three EPIC-Kitchens recordings; DirectMe, ReMEmbR, and LEDGER each build one memory over the stream and answer Q1–Q4 from it as text, without frames. Scene 4 is a separate Ego4D VQ3D clip placed on the same timeline for illustration, with memories built from that clip alone; Q5 asks for the chair's position in metres, and the map (right) shows each memory's answer. **Bottom:** the evidence each memory retrieves for Q2. LEDGER's detector labelled the object a coffee maker, but the segment description records what the wearer's hands were holding; DirectMe's graph has no node for the object, and ReMEmbR's captions near that time do not name it. Questions were selected for illustration; aggregate results are in Tables 3 and 28.

The differences follow from what each memory keeps. For Q2, "What did I take out of the cabinet?", DirectMe's graph has no node for the object, and ReMEmbR's captions near that time do not name it. Our detector labelled the object a coffee maker, but its segment description reads "a digital immersion circulator, not a coffee maker … the person's hands are holding the device", and the answerer matches it to the heating rod. For the chair in Q5, our stored position is 0.73 m from the annotated one, against 1.81 m for DirectMe and 2.37 m for ReMEmbR, which stores where the wearer was rather than where the object is.

## H ANSWER-TIME TOKEN COST

A memory is built once per recording and read many times, whereas frames are re-read at every question. This appendix measures what each memory costs the answering model at answer time, on the same questions for every arm. Tokens are the answering model's own usage, averaged per

question, and calls are language-model calls per question. Memory construction is excluded: every method builds its memory locally, with no API calls, so its cost is compute time rather than tokens. Published video models we did not run, and video-retrieval methods not run on these questions, are absent for the same reason.

Table 41: ■ Answer-time cost on HD-EPIC (2,400 questions, GPT-5.4 unless stated). Accuracy is prototype-averaged, as in Table 2. No arm here reads frames.

| Memory | Read at answer time | Acc. (%) | Input tok | Output tok | Calls |
|---|---|---|---|---|---|
| Blind | question and options only | 29.7 | 412 | 184 | 1.0 |
| ReMEmbR | captions, retrieval agent | 29.8 | 25,201 | 207 | 4.1 |
| OSNOM (adapted) | tracked object map | 30.2 | 8,762 | 111 | 1.0 |
| DirectMe | scene graph, its own prompt | 30.7 | 2,362 | 5 | 1.0 |
| LEDGER, Qwen3.5-9B | object memory | 33.6 | 1,938 | 381 | 1.0 |
| AMEGO (adapted) | hand–object interaction log | 34.7 | 792 | 124 | 1.0 |
| LEDGER | object memory | **42.6** | 1,750 | 231 | 1.0 |
| $\text{LEDGER}_{\text{agent}}$ | object memory, querying agent | 34.8 | 22,850 | 199 | 3.6 |

On HD-EPIC the stored record is read in one call of about 1.8k input tokens, and the two memories that cost less are the two that keep less: AMEGO's interaction log at 0.8k admits only handled objects, and the blind control at 0.4k stores nothing. The memories that cost more do not answer better here. ReMEmbR reads about 14 times as many input tokens over roughly four calls and scores at the blind level, because its captions carry no object positions.

Table 42: Answer-time cost on UCS-Bench (2,772 EgoLife and TeleEgo questions, GPT-5.4). Green marks arms whose answerer sees no frames, red those that do. Frame arms read 48 uniformly sampled frames up to the question time. This set has one question more than Table 3, so DirectMe's memory-only accuracy reads 29.9 here against 30.0 there.

| Evidence | Read at answer time | Acc. (%) | Input tok | Output tok | Calls |
|---|---|---|---|---|---|
| ■ Blind | question and options only | 33.8 | 452 | 149 | 1.0 |
| ■ DirectMe | scene graph | 29.9 | 1,937 | 2 | 1.0 |
| ■ LEDGER | object memory | 38.5 | 2,044 | 211 | 1.0 |
| ■ Frames | 48 frames | 44.4 | 9,615 | 50 | 1.0 |
| ■ Frames, DM routed | frames; graph on spatial q. | 42.9 | 10,406 | 49 | 1.0 |
| ■ Frames, DM own | frames and graph always | 39.8 | 11,377 | 48 | 1.0 |
| ■ Frames, LEDGER | frames; memory on spatial q. | **46.3** | 10,529 | 137 | 1.0 |

The UCS-Bench arms separate the two costs. Reading 48 frames costs about 9.6k input tokens per question, nearly five times the 2.0k that reading our record costs, and it buys 5.9 points over the record alone. Adding the record to the frames costs a further 0.9k, about a tenth of the frame budget, and recovers 1.9 points on top of frames. A memory therefore substitutes for frames at roughly a fifth of the answer-time cost, and complements them at a small fraction of it.

**Against ReMEmbR on the questions it answered.** ReMEmbR returns an answer for 2,183 of the 2,772 questions. Restricted to those, it scores 41.3% against 38.5% for our memory, so it is ahead on these single recordings, and it reads 18,917 input tokens over 3.7 calls per question against 2,041 in one call, about nine times the input for one call in place of four. The blind control scores 33.3% on the same subset. Whether that accuracy is worth the cost depends on the budget; the comparison of what each record preserves, independent of how it is read, is in Appendix E.7.