μWM: FIXED-BUDGET WORLD MEMORY FOR LONG-RUNNING WEARABLE ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wearable assistants require reliable long-horizon interaction history under strict kilobyte memory constraints, yet existing systems either lose joint spatio-temporal conditions through decoupled counters or incur unbounded storage growth in raw event logs. We introduce μWM, the first wearable world memory framework tracking joint action-, place-, and time-conditioned interaction history under a strict budget of a few tens of kilobytes. Operating locally on microcontrollers (MCUs), μWM compiles incoming semantic streams into typed state at write time, prunes duration detail via query-loss allocation while preserving counts and temporal endpoints, and resolves queries directly using deterministic readers with state-directed evidence routing. Evaluated across 14 query families, μWM achieves 86.90% accuracy at 13.38 KiB on EPIC-KITCHENS-100, outperforming equal-budget controls while using 122.6$\times$ less storage than SOTA baselines. Furthermore, it recovers 99.1% raw-log fidelity on AMB, maintains full accuracy across 32$\times$ extended ALFRED trajectories within 28.3 KiB, and fits entirely inside 64 KiB Data Tightly-Coupled Memory (DTCM) on an STM32F7 MCU at 153 to 173 mW peak power. These results establish structured state compilation as an efficient, privacy-preserving memory substrate for edge embodied intelligence.

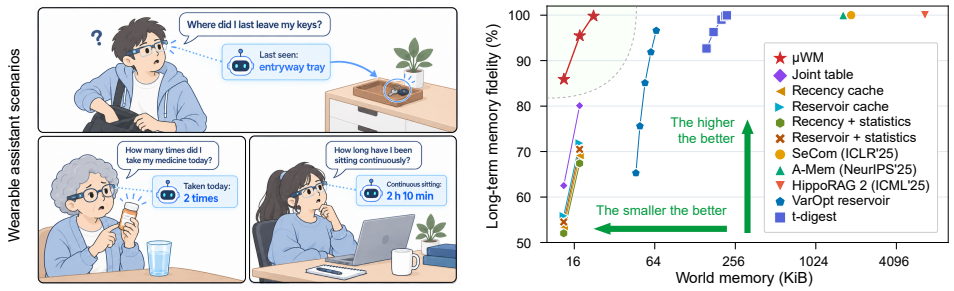


Figure 1: **Long-running wearable assistance. Left:** illustrative object-location, daily-count, and activity-duration requests. **Right:** count accuracy versus checkpoint size, where μWM maintains high fidelity within sub-16 KiB while recent memory systems require over 100$\times$ more memory.

1 INTRODUCTION

Wearable devices accompany users throughout daily life and continuously record a stream of egocentric interactions with objects and environments (Grauman et al., 2022). As illustrated in the left panel of Figure 1, a long-running assistant with persistent memory can answer routine questions, such as finding misplaced keys, tracking daily medication count, or monitoring continuous activity duration. Crucially, processing and retaining these interactions locally protects user privacy while ensuring low-latency, energy-efficient, and reliable offline operation. Furthermore, maintaining an accurate on-device world memory provides a foundation for emerging wearable AI agents and local world models to reason about long-term human context seamlessly (Fung et al., 2025).

However, maintaining such long-running memory on wearable devices is fundamentally constrained by severe hardware limits. As listed in Table 1, modern platforms offer highly restricted SRAM budgets that must be shared with sensor acquisition and wireless communication, often leaving

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.

a dedicated runtime quota of under 16 KiB. Although flash provides persistence, its coarse erase granularity and endurance limits force event buffering in RAM (Zephyr Project, n.d.; Nordic Semiconductor, n.d.), where even a modest log of one 13-byte event per second generates 366 KiB of raw data in eight hours. Existing memory compaction heuristics fail under this strict budget. FIFO and reservoir caches discard historical events, independent marginal totals destroy multi-attribute context, and joint tables degrade temporal detail. The central challenge is thus to maintain a query-adaptable world memory within a sub-16 KiB budget without sacrificing historical fidelity.

Table 1: Small working-memory budgets on wearables.

Device	Budget	Scope
Frame (Brilliant Labs, n.d.a)	256 KB	SRAM, 1 MB flash
Monocle (Brilliant Labs, n.d.b)	64 KB	SRAM, separate 64 MB FPGA RAM
PineTime (PINE64, n.d.)	64 KB	SRAM, 4 MB external flash
Bangle.js 2 (Espruino, n.d.)	256 KB	SRAM, 8 MB external flash
Watchy v3 (SQFMI, n.d.)	512 KB	SRAM, 8 MB external flash
Venu 3 (Garmin, n.d.)	64 KiB	Background-app quota

In this paper, we propose μWM, the first wearable world memory that maintains action-, place-, and time-conditioned history under a fixed budget of a few tens of kilobytes. Designed for local-first wearable architectures, μWM operates on microcontrollers (MCUs) where a lightweight perception front-end converts raw streams into semantic tuples, leaving all memory compaction and query resolution to MCU-resident code. To maximize information density within this footprint, μWM compiles incoming events at write time to incrementally track joint conditions in RAM. When memory approaches capacity, a query-loss allocation mechanism drives start-preserving demotion, dynamically shedding duration detail based on answer accuracy impact per released byte while preserving temporal endpoints and counts. For query execution, deterministic readers derive exact answers directly from the compiled state without language models, whereas complex requests leverage state-directed evidence routing to dispatch targeted data slices to a companion device only when authorized. As shown in the right panel of Figure 1, these designs allow μWM to maintain superior historical fidelity within a sub-16 KiB footprint, outperforming existing lightweight memory systems.

We evaluate μWM across 14 query families on multi-domain benchmarks, including EPIC-KITCHENS-100, AMB, and extended ALFRED trajectories averaging 13,249 events. On EPIC-KITCHENS-100, μWM achieves 86.90% overall accuracy at a 13.38 KiB checkpoint, outperforming equal-budget baselines by 11.11 percentage points on conditional queries while matching the fidelity of recent memory systems requiring over 122.6$\times$ larger storage. Furthermore, on AMB, μWM recovers 99.1% accuracy of unbounded raw logs and outperforms the strongest episodic baseline by 0.139 points, while under ALFRED trajectories extended to 32$\times$ duration, it maintains full fidelity within 28.3 KiB. Moreover, on-device profiling on an STM32F7 MCU confirms that μWM requires only 33–46 KiB of resident state, fitting entirely inside the 64 KiB zero-wait DTCM to eliminate per-event flash writes with a peak power of only 153–173 mW.

To summarize, our work makes the following key contributions:

- We introduce μWM, the first wearable world memory that maintains action-, place-, and time-conditioned history under a target sub-16 KiB budget, establishing a privacy-preserving, local-first paradigm for on-device agents with continuous contextual awareness.
- We propose a compact memory architecture that unifies write-time compilation over joint temporal relations with query-directed precision allocation. This design leverages start-preserving demotion to reclaim storage without altering count fidelity, integrates state-directed evidence routing for complex queries, and provides theoretical lower bounds on temporal retention.
- We conduct extensive multi-domain evaluations and on-device profiling to demonstrate that μWM maintains high fidelity over long event streams while fitting entirely within MCU-resident SRAM, decisively outperforming existing memory systems in storage efficiency.

2 μWM DESIGN

As illustrated in Figure 2, μWM operates across three coordinated stages to compile continuous interaction streams into a compact memory state using either query-directed precision allocation or tiered compaction. The compiled state powers two distinct query resolution pathways. Routine requests for daily habit counts, activity timestamps, and locations are derived locally from typed reads without on-device language models. Conversely, complex open-ended requests use typed

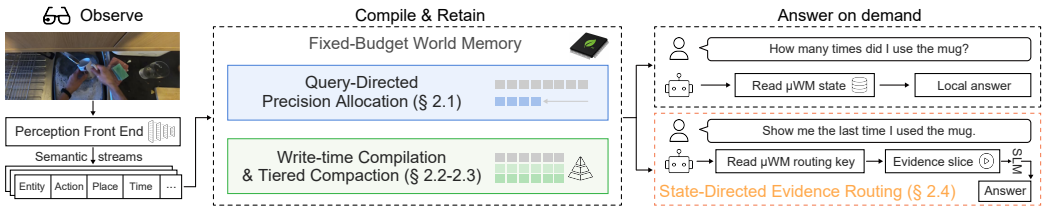


Figure 2: μWM overview. Semantic events are compiled into MCU-resident memory, enabling instant local reads for routine queries and selective evidence routing for open-ended requests.

reads to establish entity-time anchors, retrieving targeted evidence slices from an onboard archive and routing them to a companion model only under explicit user authorization.

Representation objective. For a target query family $\mathcal{Q}$, μWM updates its memory state upon event arrival and evaluates queries $q \in \mathcal{Q}$ by approximating full-history answers F_q within a packed budget B,

$$S_t = U(S_{t-1}, e_t), \qquad \text{bytes}_{\text{packed}}(S_t) \le B, \qquad R_q(S_t) \approx F_q(e_{1:t}).$$

The state assigns exact scalar fields or bounded temporal summaries based on query demands to maximize historical fidelity under the budget.

Event interface. The input is an ordered stream of semantic tuples $e_t = (n_t, v_t, a_t, b_t, \ell_t)$, where n_t denotes the entity ID, v_t the action label, $[a_t, b_t]$ the temporal interval, and ℓ_t the place ID mapped to visual codebook centroids. The online update $S_t = U(S_{t-1}, e_t)$ incrementally compiles each event into entity, temporal, relational, and spatial representations, updating named state fields directly rather than appending raw feature vectors.

2.1 Query-Directed Precision Allocation

Joint-condition records. To preserve multi-attribute conditional context, the adaptive layout organizes temporal support by entity. Each record retains joint action and place labels, enabling queries to filter by individual attributes or their conjunction. A full singleton record stores action, place, start, and end timestamps in 11 bytes. A folded record consolidates repeated occurrences of the same action and place into 19 bytes, storing joint labels, total count, first and last start endpoints, and cumulative duration. For counts of two or more, the count reader assigns one count unit to each stored start endpoint and distributes the remaining count uniformly between them.

Start-preserving demotion. When memory budget pressure increases, start-preserving demotion reclaims storage by reconstructing event durations through a global decoder while strictly preserving joint labels, total counts, and start endpoints. Singleton and folded records shrink from 11 to 7 bytes and from 19 to 15 bytes, respectively, freeing 4 bytes per record without altering represented conditional count accuracy. The decoder uses a fixed mean duration of 3.1 seconds stored in a 10-byte header. Because duration-based reads are affected by demotion, the allocation policy dynamically weighs demotion against temporal merging and record deletion.

Answer-aware byte allocation. Upon memory overflow, the updater evaluates duration demotion, same-label temporal merging, and record deletion. It iteratively selects the candidate operation c^* that minimizes query-loss impact per released byte until the state fits the budget,

$$c^* = \arg\min_{c \in \mathcal{C}(S)} \frac{\widehat{\mathcal{L}}_{\mathcal{Q}}(S, c(S))}{\text{bytes}(S) - \text{bytes}(c(S))}.$$

Candidate costs are calculated from the current state and probes across target query families. For conditional counts, the loss formulation incorporates a normalized cumulative count discrepancy,

$$\frac{1}{|I|} \int_I \min\big(1, \big|F_S(t) - F_{c(S)}(t)\big|\big)\, dt,$$

where F represents the cumulative count trajectory over time span I. This integral measures changes to the represented answer trajectory, corresponding to the expected disagreement in rounded counts under uniform cutoff offsets. Additional terms track activity-bin discrepancies, changes to primary location or entity associations, and lost scalar duration mass.

Lossless-prefix storage and update procedure. Incoming events are held in a causal compressed prefix with a 12-byte header before triggering state compaction. When memory limits are reached, this prefix is compiled into bounded state, preserving short-term event histories exactly. Each new observation updates scalar statistics and appends a full-precision record. When memory exceeds the cap, query-loss allocation triggers demotion, merging, or deletion. Merging combines records with matching action and place labels, converting three 7-byte singletons into a single 15-byte folded record to achieve a net saving of 6 bytes. Appendix A.1 provides the detailed update procedure.

Precision policies and budget settings. We evaluate two precision operating modes under query-directed allocation. Adaptive precision applies duration demotion selectively when chosen by the query-loss rule, preserving measured durations for duration-sensitive queries. Fixed precision applies the duration decoder to every record upon entry into typed state, maximizing storage allocated to count and timestamp histories. As shown in the right panel of Figure 1, fixed precision achieves high count fidelity across 16, 24, and 48 KiB core caps. Table 2 compares both precision policies at a 16 KiB budget, Table 3 evaluates performance across 4, 8, and 16 KiB budgets, and Appendix A.3 details the byte accounting for tiered reference layouts.

2.2 Tiered State Compaction

Tiered state compaction maintains detailed statistics for active entities while compressing long-tail history into coarse summaries. This strategy uses a fixed layout with constant capacities across active entities A, dormant entities D, place visits V, and a recent partner window W.

Active and dormant entity tiers. Memory is partitioned based on downstream query access patterns. Active slots in A store comprehensive statistics, including bounded action rows, weighted temporal intervals, and place associations for the current working set. When capacity is reached, the least-recently used active record moves to dormant storage in D, retaining its entity ID, first seen timestamp, and total count. As demonstrated in Figure 3, standard LRU eviction discards history completely, forcing an entity counter to reset to 1 upon re-entry. In contrast, μWM restores dormant historical state into A upon entity reappearance, preserving long-tail continuity across active slot turnover.

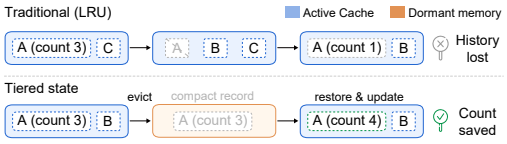


Figure 3: History across eviction. Dormant storage preserves entity counts when conventional LRU replacement would reset them.

Spatial, relational, and temporal representations. Place table V maintains first and latest timestamps alongside visit intervals, extending visits when events occur within 60 seconds. Recent window W retains co-occurring entity pairs from the past 30 seconds to update symmetric relational counts. To bound temporal representation size, active entity intervals are summarized via a weighted temporal coreset. When interval storage reaches capacity, adjacent interval pairs with the smallest midpoint gap are merged through weighted averaging,

$$\left(\frac{w_i a_i + w_{i+1} a_{i+1}}{w_i + w_{i+1}}, \frac{w_i b_i + w_{i+1} b_{i+1}}{w_i + w_{i+1}}, w_i + w_{i+1} \right),$$

retaining exact total event counts while strictly bounding detailed interval timing precision.

Write-time compilation and static footprint. Every event is compiled in place without buffering raw logs. Update operations adjust scalar counters and maintain table capacities in $O(M)$ worst-case time. For entity capacity M and place capacity K, packed state memory follows

$$B_{\text{stream}} = 16 + 453M + 77K \quad \text{bytes.} \tag{1}$$

Appendix A.3 details field layout specifications, and Appendix B provides update invariants.

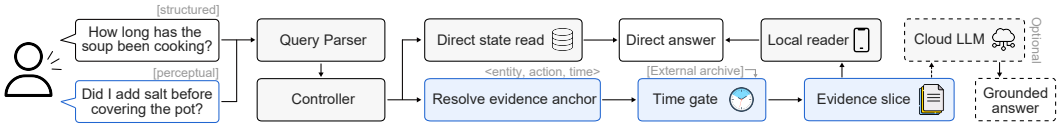


Figure 4: State-directed evidence routing workflow. The Query Parser and Controller resolve requests through direct state reads or entity-time temporal gating to guide external evidence retrieval.

2.3 Algebraic Sufficiency and Theoretical Boundaries

Scalar sufficiency. To guarantee uncompromised answer fidelity over arbitrarily long event streams, μWM models scalar statistics using closed-form streaming reductions. Routine queries regarding counts, temporal extremes, and spatial associations reduce to commutative and associative binary operations over the incoming stream. By maintaining exact algebraic accumulators in place during event ingestion, μWM computes precise scalar reductions without full log re-scans.

Proposition 1 (Fixed-memory scalar sufficiency)**.** *For fixed capacities and field widths, the memory allocation in Equation 1 resolves all supported scalar reductions identically to a full scan of the entire event stream, provided its accumulators do not overflow.*

Counting beyond event buffers. Conventional episodic memory baselines store unmodified raw input tuples in fixed buffers, causing rapid state blowup or premature event eviction over extended temporal horizons. We formalize why these tuple-centric designs fail under tight budget constraints by establishing the fundamental theoretical lower bound on required tuple-buffer memory.

Proposition 2 (Exact counting beyond an event buffer)**.** *A memory storing at most M unmodified input tuples without counters can count streams of length up to H exactly only if $M \geq H$. A scalar counter requires only $\lceil \log_2(H+1) \rceil$ bits.*

This lower bound demonstrates that stateful algebraic accumulators achieve exponential space compression over raw tuple buffers, compressing an arbitrary unbounded event stream of length H into merely $\lceil \log_2(H+1) \rceil$ bits. This crucial distinction forms the theoretical foundation for maintaining continuous exact count history within a default sub-16 KiB MCU SRAM budget.

Temporal representation boundary. While scalar statistics can be preserved with logarithmic memory growth, answering arbitrary temporal queries over continuous event histories introduces strict information-theoretic boundaries.

Proposition 3 (Exact temporal queries require growing memory)**.** *Answering exact interval occupancy over H discrete time bins requires at least H bits in the worst case.*

Because exact arbitrary interval occupancy requires linear memory scaling with respect to temporal horizon H, fixed-budget architectures cannot preserve timestamp occupancy indefinitely. This fundamental limit directly justifies μWM's hybrid architecture, which pairs exact scalar accumulators with bounded weighted temporal coresets to maintain timing fidelity under tight budget caps. Appendix B provides full proofs, and Appendix B.1 specifies the algebraic state reductions.

2.4 State-Directed Evidence Routing

μWM transforms its compact resident state into a routing index, dispatching complex open-ended requests to targeted external evidence slices through resolved entity and temporal anchors.

Routing architecture. As shown in Figure 4, the Query Parser maps input requests to typed queries q. The Controller executes direct state reads for structured queries or produces a routing descriptor when evidence is required,

$$\mathrm{Query}(q, S_T) \longrightarrow \langle \widehat{y}, \mathcal{T}_{\mathrm{bound}} \rangle,$$

where $\widehat{y}$ is a local answer or $\perp$, and $\mathcal{T}_{\mathrm{bound}}$ contains entity, action, and temporal boundaries.

Anchor-guided retrieval. For requests requiring rich perceptual evidence, the temporal anchor $\tau(q, S_T)$ and search radius Δ define a temporal gate,

$$\mathcal{C}(q) = \{j \mid |t_j - \tau(q, S_T)| \leq \Delta\}, \qquad \widehat{a} = r(q', \mathcal{E}_{\mathcal{C}(q)}).$$

Table 2: End-to-end retrieval fidelity comparison between different memory management methods. The zlib6 column applies an identical lossless codec to each World Memory checkpoint. Parentheses give text-checkpoint caps in KiB.

Method	World Memory (KiB)↓	zlib6 (KiB)↓	Nonzero counts (%)↑	All 14 families (%)↑
HippoRAG 2	6697.87	4417.77	100.00	100.00
SeCom	1879.88	313.09	100.00	100.00
A-Mem	1640.46	337.59	100.00	100.00
Text FIFO (512)	354.02	40.79	73.62	87.35
Text FIFO (256)	211.39	24.63	40.20	65.03
Text uniform (128)	112.30	13.90	19.76	49.11
Text FIFO (64)	58.29	7.15	19.61	36.01
Joint table, 4 cells	13.35	8.77	55.35	82.14
Joint table, 2 cells	13.35	8.78	55.35	82.14
Typed recency cache	13.35	8.62	49.46	75.00
Typed recency + stats	13.35	8.44	48.24	75.60
μWM (Query-directed)	13.38	8.99	73.39	86.90
μWM (Fixed-precision)	13.38	8.94	73.51	87.65

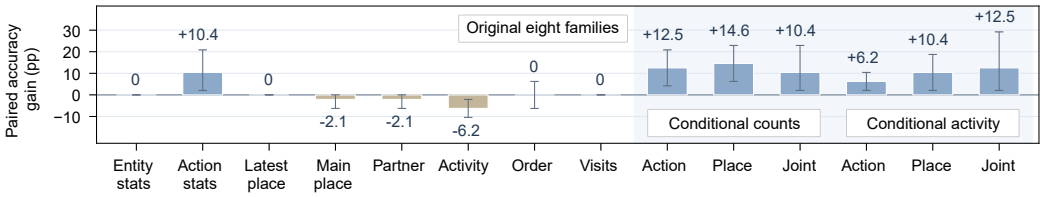


Figure 5: Conditional-history gains. Adaptive μWM minus the two-cell joint table at 16 KiB across fourteen families, with paired 95% participant intervals.

This search gate restricts external retrieval to timestamped evidence $\mathcal{E}_{\mathcal{C}(q)}$. The retrieved slice feeds a downstream companion reader under explicit user authorization. By maintaining index anchors on the MCU while keeping raw archives separate, μWM handles complex perceptual requests while preserving a strict resident memory footprint. Appendix F.4 provides archive and retrieval details.

3 Experiments

3.1 Experimental Setup

Datasets and query suites. We evaluate μWM across three multi-domain benchmarks spanning varied temporal horizons and interaction complexities under a strictly unified protocol where semantic events are ingested without prior knowledge of query arrivals. First, an EPIC-KITCHENS-100 suite evaluates continuous multimodal streams across 21,594 events from 96 egocentric recordings across six participants (Damen et al., 2022), examining both a 672-question shared-model benchmark and a full catalog of 66,107 conditional-count queries. Second, the AMB egocentric benchmark supplies 8729 confirmatory questions over 31 videos covering eight high-level interaction query types (Goletto et al., 2024). Third, ALFRED household trajectories extend stream history up to 32$\times$ duration with an average horizon of 13249 events across 105 trajectories (Shridhar et al., 2020). A disjoint 506-trajectory validation split tests state-directed evidence routing. Appendices B.1, C, and F detail query algebra, dataset splits, and trajectory protocols.

Baselines and shared readout. As summarized in Table 2, we benchmark μWM against recent memory architectures including SeCom, A-Mem, and HippoRAG 2 (Pan et al., 2025; Xu et al., 2025; Jiménez Gutiérrez et al., 2025), alongside sliding-window text selectors, typed recency caches, and streaming estimators such as VarOpt, reservoir sampling, and double-precision t-digest (Vitter, 1985). To isolate memory representation fidelity from language model variance, all shared-model evaluations route retrieved facts through identical deterministic statistical readers into a frozen downstream Qwen2.5-VL-7B-Instruct model operating in text-only mode with greedy decoding.

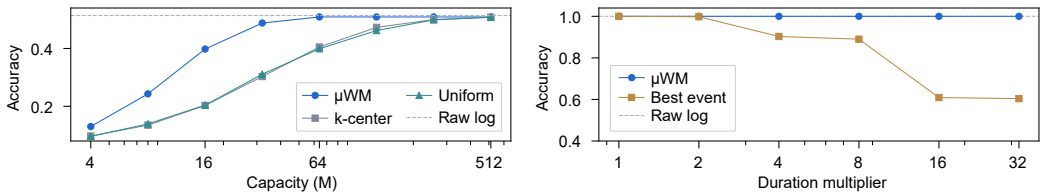


Figure 6: Record density and long-history retention across continuous streams. Left, AMB answer accuracy at matched record counts. Right, ALFRED historical fidelity at fixed memory capacity.

Prompting structures, contextual framing, and output token budgets are strictly identical across all evaluated memory methods. Appendix C provides the full context codecs and serialization layouts.

Budgets and metrics. All compact typed architectures operate under a strict 16 KiB core memory cap. Reported checkpoint allocations strictly account for all persistent components including state fields, semantic labels, decoders, indices, and framing metadata, while model weights and transient workspaces remain separate. Performance is scored via exact-match accuracy, with parse failures and unanswered questions penalized as errors. Statistical confidence intervals at 95% confidence are calculated using 10,000 bootstrap resamples across participants for EPIC-KITCHENS-100 and video clusters for AMB. Appendix A.3 specifies complete memory accounting rules.

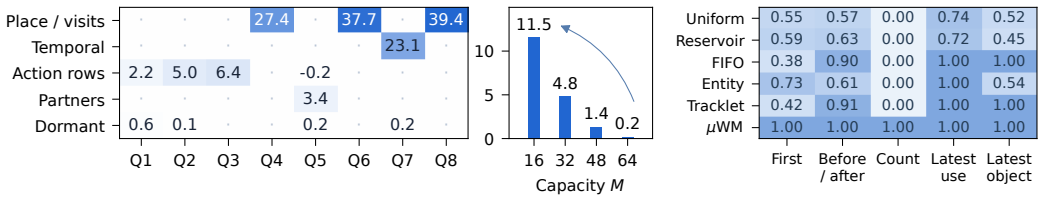
3.2 Main Results

We empirically evaluate end-to-end query retrieval accuracy, long-horizon temporal stream retention, and record storage efficiency across three diverse multimodal interaction benchmarks.

Shared-model retrieval accuracy and storage efficiency. As summarized in Table 2, μWM delivers competitive retrieval fidelity while reducing storage footprints by orders of magnitude compared to published memory architectures. Operating within an average checkpoint of only 13.38 KiB, query-directed μWM reaches 86.90% accuracy across all 14 query families. Compared to the equal-budget joint table at 13.35 KiB, μWM achieves a paired gain of 4.76 percentage points with a 95% interval of $[1.93, 7.59]$, boosting nonzero conditional-count accuracy from 55.35% to 73.39%. Fixed-precision μWM matches this performance at 87.65% accuracy by reallocating freed duration bytes to expand count and endpoint capacity. In contrast, published graph and episodic baselines such as SeCom, A-Mem, and HippoRAG 2 require 1640.46 to 6697.87 KiB on average, representing a storage overhead exceeding 122.6 times that of μWM. Under an identical zlib6 codec, μWM occupies merely 8.99 KiB, compared to 313.09 to 4417.77 KiB for published systems. Furthermore, sliding-window text selectors require 512 KiB caps to approach comparable accuracy, demonstrating that unstructured event logs waste capacity on redundant tokens.

Conditional history gains and count frontier. Figure 5 details performance gains across individual query families over the equal-budget joint table. The accuracy advantage of μWM concentrates on queries requiring joint semantic and temporal history. On the six conditional families covering action-, place-, and jointly-conditioned counts and activity timing, accuracy increases from 73.96% to 85.07%, representing a decisive paired gain of 11.11 percentage points $[4.17, 18.06]$. Furthermore, as demonstrated in the count frontier in Figure 1, fixed-precision μWM reaches 95.43% count accuracy at 17.58 KiB, outperforming the joint table baseline by 15.34 points $[4.23, 27.33]$. To match this fidelity, event-sampling baselines require substantial memory overheads, with VarOpt requiring 65.65 KiB and t-digest requiring 155.87 KiB, corresponding to 3.7 and 8.9 times the storage footprint of μWM. This confirms that directly compiling count trajectories is fundamentally more space-efficient than reconstructing conditional history from sampled event streams.

Real egocentric memory performance and long-horizon resilience. On the AMB egocentric benchmark, μWM achieves 0.302 accuracy using predicted entity identities and 0.509 with given identities across 31 held-out videos. On the 8-type confirmatory protocol, μWM achieves 0.509 accuracy, recovering 99.1% of unbounded raw-log fidelity while outperforming the strongest 64-event control at 0.405. As illustrated in the left panel of Figure 6, μWM achieves high record density at capacity $M{=}64$, whereas uniform and k-center sampling require an $8\times$ capacity expansion to $M{=}512$ to match comparable fidelity. Over the full 100-video benchmark, μWM scores 0.569



(a) Field ablations (b) Dormant tier (c) Independent history fields

Figure 7: Complementary memory mechanisms. Accuracy drops after field removal at M=64 (a) and dormant-tier removal (b). Accuracy at 32× ALFRED (c). Dots in (a) denote zero.

against 0.430 for the best 64-event control and 0.590 for the full raw log. Finally, on ALFRED long-horizon streams evaluated in the right panel of Figure 6, μWM maintains an exact accuracy of 1.0 across stream replays extending up to 32× (13249 mean events) within a constant 28.3 KiB footprint, matching the unbounded raw event log throughout. While event-sampling buffers experience overflow and collapse to 0.604, compiled state preserves exact historical fidelity without storage growth. Appendices C, D, and D.4 provide detailed baseline comparisons and per-type breakdowns.

3.3 Ablation Studies

We isolate the contribution of each architectural component across budget allocations, memory fields, compaction tiers, and noisy perceptual inputs.

Query-directed allocation across budgets. Table 3 evaluates query-directed allocation against the best complete-query control across increasing memory caps. While performance gains are visible at 4 KiB, the margin widens significantly as memory expands. At 8 KiB, nonzero conditional-count accuracy rises from 35.36% to 53.20%, representing a +17.85 percentage point gain. At 16 KiB, this advantage reaches +20.96 points, increasing from 62.62% to 83.58%. Expanding byte capacity allows query-directed allocation to dedicate additional accumulators to high-order conditional counts, converting extra storage directly into answer precision. Appendix C.2 provides full confidence intervals.

Table 3: Query-directed allocation across budgets. Accuracy (%) and count gains over the best complete-query control.

Cap	All 14 families		Nonzero counts		
(KiB)	μWM	Ctrl.	μWM	Ctrl.	Gain
4	57.04	55.49	26.88	22.92	+3.96
8	72.08	68.66	53.20	35.36	+17.85
16	88.13	83.70	83.58	62.62	+20.96

Typed fields and temporal compression. Figure 7(a) isolates individual component contributions by measuring accuracy drops upon field removal at active capacity $M = 64$. Removing dedicated place and visit tables severely degrades last location (Q4), use location (Q6), and visit time (Q8) queries by 27.4, 37.7, and 39.4 percentage points, respectively. Removing weighted temporal coresets lowers use time accuracy (Q7) by 23.1 points. Crucially, retaining just twelve weighted intervals achieves 0.767 accuracy on Q7, closely approaching unbounded raw uncompressed intervals (0.770) while vastly outperforming coarse endpoint-only baselines (0.535). Action rows and symmetric partner maps supply targeted gains for relative temporal order (Q3, up to 6.4 points) and co-use relationships (Q5, 3.4 points). Appendices E.1 and E.3 provide extended field breakdowns.

Dormant compaction and history decoupling. Dormant tier compaction safeguards long-tail history under tight slot budgets (Figure 7(b)). Under high eviction pressure at $M = 16$, removing dormant summaries causes an 11.5 percentage point drop in accuracy, whereas the penalty tapers to 0.2 points at $M = 64$ as active slots fully cover the working set. Furthermore, Figure 7(c) demonstrates that decoupling history into dedicated scalar fields enables μWM to maintain 1.000 accuracy across all query types over 32× ALFRED durations without buffer contention. In sharp contrast, all event-sampling baselines collapse to zero on exact counts, while FIFO and tracklet buffers drop to 0.383 and 0.420 on ordering queries. This empirical gap directly validates the count preservation boundary in Proposition 2.

Robustness to predicted identities. To evaluate resilience against upstream visual perception noise, we test μWM using video-disjoint noun predictions at 0.340 event accuracy (Appendix D.6). μWM reaches 0.302 accuracy compared to 0.246 for the matched event memory baseline, yielding a

$+0.056$ point gain $[+0.041, +0.070]$ across 8729 questions. This gain remains consistently positive across interaction query families Q1 through Q7, confirming that structured compilation remains robust when fed noisy predicted labels rather than ground-truth annotations.

3.4 On-Device Profiling

We profile three μWM configurations on an embedded ARM Cortex-M7 MCU (STM32F746G running at 216 MHz) to evaluate physical hardware execution (STMicroelectronics, 2025). Table 4 reports resident state footprints, SRAM utilization, and empirical peak power draw across a regulated 3.30 V supply rail. Across all configurations, resident memory remains between 33.02 KiB and 45.64 KiB, consuming under 15% of total system SRAM, preserving substantial capacity for real-time sensing drivers. Crucially, the entire active state fits inside the chip's 64 KiB DTCM, ensuring single-cycle zero-wait-state access without bus contention or pipeline stall jitter. State updates execute deterministically in RAM with $O(M)$ worst-case latency, completely avoiding destructive per-event flash programming wear while consistently sustaining ultra-low active power consumption between 153 mW and 173 mW.

Table 4: On-device profiling of μWM.

μWM Config.	RAM (KiB)	SRAM (%)	Peak power (mW)
ALFRED	33.02	10.32	$\sim$153
AMB	33.64	10.51	$\sim$157
AMB + codebook	45.64	14.26	$\sim$173

4 Related Work

Episodic video and agent memory architectures. Early video QA established constant-size representations (Bärmann & Waibel, 2022), while modern video assistants organize multimodal evidence for long-horizon retrieval (Wang et al., 2024; Song et al., 2024; Ren et al., 2024; Diko et al., 2025; Wang et al., 2025; Zhang et al., 2024; Goletto et al., 2024; Fan et al., 2024; Huang et al., 2025). Embodied agents and memory frameworks extend localization and structured consolidation (Manigrasso et al., 2026; Fan et al., 2025; Yang et al., 2025a; Das & Roy, 2026), while agent memories and world models leverage graphs, hierarchical retrieval, or latent updates (Pan et al., 2025; Xu et al., 2025; Jiménez Gutiérrez et al., 2025; Sarthi et al., 2024; Asai et al., 2024; Jiménez Gutiérrez et al., 2024; Ha & Schmidhuber, 2018; Hafner et al., 2023; Hansen et al., 2024; Bruce et al., 2024). However, these approaches either incur unbounded storage growth or fail on exact cumulative queries. μWM bridges this gap by compiling history under a strict byte cap, directly answering cumulative questions while routing descriptive lookups to evidence stores.

Bounded neural context and streaming summaries. Attention-cache compression (He et al., 2024; Zhang et al., 2025; Kim et al., 2025; Zeng et al., 2025; Xiao et al., 2024; Zhang et al., 2023) and soft-memory banks (Mu et al., 2023; Ge et al., 2024; Wang et al., 2023) constrain neural overhead during extended context processing, whereas recurrent state-space models encode streams via structured updates but bottleneck on exact, multi-step recall (Gu et al., 2022; Dao & Gu, 2024; Yang et al., 2024; 2025b; Sun et al., 2025; Arora et al., 2024a;b). Alternatively, classic streaming sketches, coresets, and quantile estimators maintain summary statistics for frequency or distributional metrics (Arasu et al., 2006; Cormode & Muthukrishnan, 2005; Metwally et al., 2005; Gonzalez, 1985; Hochbaum & Shmoys, 1985; Cohen et al., 2009; Dunning & Ertl, 2019), yet lack joint temporal-semantic conditioning for complex queries. μWM unifies streaming algorithms with memory representations via query-directed precision allocation, preserving joint temporal and semantic support while pruning uninformative details within a minimal byte footprint.

5 Conclusion

We present μWM, the first wearable world memory framework tracking joint action-, place-, and time-conditioned history under a strict kilobyte budget. μWM compiles streams into joint conditions at write time, prunes duration via query-loss allocation, and resolves queries using deterministic readers with evidence routing. On EPIC-KITCHENS-100, it achieves 86.90% accuracy at 13.38 KiB, outperforming equal-budget controls, while using 122.6$\times$ less storage than baselines. It recovers 99.1% AMB fidelity, preserves 32$\times$ ALFRED trajectories within 28.3 KiB, and fits the 64 KiB DTCM of an STM32F7 MCU, delivering an efficient substrate for edge embodied agents.

AI USE STATEMENT

Generative AI tools were used to assist with mathematical formulation and proof development, the design of experiments and diagnostic analyses, and code implementation and debugging. The core technical design of this work, and all manuscript drafting and editing, were completed by the authors. The authors take responsibility for the final content of this work.

REFERENCES

Arvind Arasu, Shivnath Babu, and Jennifer Widom. The CQL continuous query language: Semantic foundations and query execution. *The VLDB Journal*, 15(2):121–142, 2006.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/448fc91f669c15d10364ee01d512cc10-Abstract-Conference.html.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning (ICML)*, 2024b. URL https://proceedings.mlr.press/v235/arora24a.html.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Leonard Bärmann and Alex Waibel. Where did I leave my keys? – Episodic-Memory-Based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. doi: 10.1109/CVPRW56347.2022.00162.

Brilliant Labs. Frame hardware manual, n.d.a. URL https://docs.brilliant.xyz/frame/hardware/. Accessed September 23, 2026.

Brilliant Labs. Monocle hardware manual, n.d.b. URL https://docs.brilliant.xyz/monocle/hardware/. Accessed September 23, 2026.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Edith Cohen, Nick Duffield, Haim Kaplan, Carsten Lund, and Mikkel Thorup. Stream sampling for variance-optimal estimation of subset sums. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009. URL https://arxiv.org/abs/0803.0473.

Graham Cormode and S Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, 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.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Kousik Das and Debaditya Roy. Imprint: Online memory compression for long-horizon egocentric QA. *arXiv preprint arXiv:2607.00696*, 2026. URL https://arxiv.org/abs/2607.00696.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. ReWind: Understanding long videos with instructed learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13734–13743, 2025.

Ted Dunning and Otmar Ertl. Computing extremely accurate quantiles using t-Digests. *arXiv preprint arXiv:1902.04023*, 2019. URL https://arxiv.org/abs/1902.04023.

Espruino. Bangle.js 2, n.d. URL https://www.espruino.com/Bangle.js2. Accessed September 23, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision (ECCV)*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6342–6352, 2025.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, Paden Tomasello, and Jitendra Malik. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

Garmin. Venu 3, n.d. URL https://developer.garmin.com/connect-iq/articles/device-reference/venu3.html. Connect IQ device reference. Accessed September 23, 2026.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b276510ec2d3f6613a8b60c41ff0438-Abstract-Conference.html.

Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. AMEGO: Active memory from long EGOcentric videos. In *European Conference on Computer Vision (ECCV)*, 2024.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, 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 (CVPR)*, pp. 18995–19012, 2022.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cf73d57b6dcda32b293df7c2d5341f49-Abstract-Conference.html.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, Ning Zhang, Yong Jae Lee, and Miao Liu. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning (ICML)*, pp. 21497–21515, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. InfiniPot-V: Memory-constrained KV cache compression for streaming video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/caef5f5e658aa1f7565f063a2cd99726-Abstract-Conference.html.

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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3951–3960, 2026.

Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Efficient computation of frequent and top-k elements in data streams. In *International Conference on Database Theory (ICDT)*, 2005.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3d77c6dcc7f143aa2154e7f4d5e22d68-Abstract.html.

Nordic Semiconductor. NVMC: Non-volatile memory controller, n.d. URL https://docs.nordicsemi.com/r/bundle/ps_nrf52840/page/nvmc.html. nRF52840 Product Specification. Accessed September 23, 2026.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. SeCom: On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e56f394bbd4f0ec81393d767caa5a31b-Abstract-Conference.html.

PINE64. PineTime (single page), n.d. URL https://pine64.org/documentation/PineTime/_full/. Accessed September 23, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8a2acd174940dbca361a6398a4f9df91-Abstract-Conference.html.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

SQFMI. Hardware, n.d. URL https://watchy.sqfmi.com/docs/hardware/. Watchy documentation, Revisions. Accessed September 23, 2026.

STMicroelectronics. *STM32F745xx and STM32F746xx Datasheet*. STMicroelectronics, July 2025. URL https://www.st.com/resource/en/datasheet/stm32f746ng.pdf. Revision 5, embedded SRAM and Cortex-M7 specifications.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985. doi: 10.1145/3147.3165.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ebd82705f44793b6f9ade5a669d0f0bf-Abstract-Conference.html.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d13a3eae72366e61dfdc7eea82eeb685-Abstract-Conference.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4904fad153f6434a7bcf04465d4be2cc-Abstract-Conference.html.

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

Zephyr Project. Non-volatile storage (NVS), n.d. URL https://docs.zephyrproject.org/latest/services/storage/nvs/nvs.html. Accessed September 23, 2026.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

APPENDIX CONTENTS

A State architecture and implementation **16**
A.1 Update procedures and event inputs 16
A.2 Active epochs and relational updates 17
A.3 Field widths and the serialized contract 17
A.4 Record capacity and the storage crossover 18

B Query algebra and formal guarantees **18**
B.1 Fixed memory and scalar reads 18
B.2 Counting and temporal information requirements 19

C Evaluation protocol and baselines **19**
C.1 Data interfaces and frozen choices 19
C.2 Controls, pairing, and uncertainty 20
C.3 Shared-model memory comparison 21
C.4 Pixel-to-answer reference systems 22

D Complete AMB results and robustness **23**
D.1 The complete capacity frontier 23
D.2 Query-family transfer to the complete benchmark 23
D.3 Video weighting and leave-one-out sensitivity 24
D.4 Answer preservation after compilation 24
D.5 Recognition and design-only prior selection 25
D.6 Paired gains under the noisy channel 25

E Field mechanisms and temporal capacity **26**
E.1 Field contributions and query dependencies 26
E.2 Dormant-tier history coverage 27
E.3 Temporal design and disjoint confirmation 27

F Household transfer and long-history scaling **27**
F.1 Transfer on natural household trajectories 27
F.2 Equal-byte scaling across every selector 28
F.3 Early, cumulative, and recent-history retention 28
F.4 State-directed routing for description retrieval 29

G System efficiency and reproducibility **30**
G.1 Latency and serialized growth 30
G.2 Resident-history and working-set checks 31

A STATE ARCHITECTURE AND IMPLEMENTATION

A.1 UPDATE PROCEDURES AND EVENT INPUTS

Algorithm 1 gives the variable-precision write path from Section 2.1. Algorithm 2 specifies the tiered state compaction of Section 2.2. The former enforces a byte cap through precision and retention choices. The latter maintains active and dormant tiers within a fixed layout. Both process semantic inputs online.

Algorithm 1 Query-directed allocation for one arriving interaction

Require: Event e, state S, total budget B, frozen decoder D
if S is a compressed prefix **then**
  Encode the prefix followed by e into candidate packet P
  if $|P| + |D| \leq B$ **then**
    return P, including its 12-byte mode header
  end if
  Replay the existing prefix into an empty bounded typed state
end if
Archive a completed place visit and compact its reserved history share
Update entity scalars, current visits, and the bounded partner window
Add one full-precision support with its action, place, and interval
while $\mathrm{bytes}_{\mathrm{packed}}(S) + |D| > B$ **do**
  Enumerate byte-releasing demotions, same-label merges, and deletions
  if a candidate exists **then**
    Apply the candidate with the smallest query-loss cost per released byte
  else
    Evict the least-recent entity row whose optional supports are empty
  end if
end while
Release temporary candidate lists and query histograms
return The packed typed state, including all module and mode headers

Algorithm 2 Tiered state compaction for one interaction

Require: State $S = (A, D, V, W)$, event $e = (n, v, a, b, \ell)$, capacities M, K
if $n \notin A$ **then**
  Initialize $A[n]$, restoring first time and count from $D[n]$ if present
  if $|A| > M$ **then**
    Move the least-recent other active entity's coarse fields to D
    Keep at most M dormant rows, evicting the least-recent old row
  end if
end if
Increment entity count, set latest time and action, and update the action row
Insert $(a, b, 1)$ and merge an adjacent pair if temporal capacity is exceeded
Assign latest place and increment its capped place histogram
Filter W to end times $> a - 30$ and keep its latest $M - 1$ entries
for $(m, b_m) \in W$ with $m \neq n$ **do**
  Add weight 2 if $b_m \geq a$, else 1, to resident symmetric partner maps
end for
Append (n, b) to W and extend or append the place visit in $V[\ell]$
Restore all nested capacities and **return** S

Entity rotation precedes statistics and partner updates, so both use the post-rotation active set. Entity and action endpoints refer to event starts. The partner window uses end times to detect overlap, and place visits use both bounds. Figure 8 shows actual input frames with their annotated actions, intervals, and frozen place assignments.

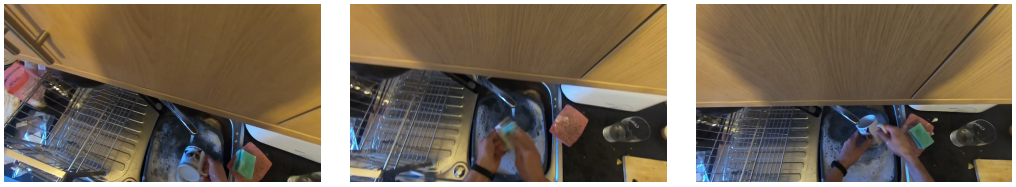


Figure 8: Scene-to-event inputs. EPIC-KITCHENS video P30_111 supplies the mug interactions, annotated intervals, and frozen visual-place IDs.

A.2 Active epochs and relational updates

An active epoch lasts while an entity remains in the rich table. Table 5 shows which fields survive dormant storage and which restart on reactivation.

Table 5: Field lifetime under active and dormant rotation.

Field	While active	After dormant reactivation
Entity first time	Preserve earliest start	Restore preserved first time
Entity count	Increment each arrival	Restore, then increment
Entity latest time	Set current start	Set from new arrival
Action-specific rows	Accumulate resident actions	Start a new detailed epoch
Place and partner counts	Update capped maps	Start a new detailed epoch
Temporal representatives	Merge epoch intervals	Start with the new interval
Place visits	Maintain separately	Unchanged by entity rotation

The partner window retains end times greater than $a_t - 30$ and at most the latest $M - 1$ inserted entries before appending the new event. An overlapping interaction contributes weight 2 and a recent non-overlapping interaction contributes weight 1. Both partner maps receive the contribution when both entities are active. Separate recent records represent separate interactions. The record cap bounds the scan to $O(M)$, including arbitrarily dense arrivals, and a serialized diagnostic flag records truncation.

Local replacement follows the statistic. Action maps remove the least-recent action at 16 rows. Place histograms remove a minimum-count key at four rows. Partner maps also use a minimum-count key, but reject a new contribution whose weight is smaller than that minimum. Ties follow deterministic iteration order. Each place keeps its eight latest visit intervals. Appendix E tests the effects of these field capacities.

A.3 Field widths and the serialized contract

The tiered reference uses a fixed layout of little-endian fixed-width records and pads each collection to capacity. Length fields and reserved IDs distinguish unused storage from valid records. An active row comprises a 21-byte header, 224 action bytes, 144 temporal bytes, 20 place-count bytes, and 24 partner bytes, totaling 433. Its paired dormant row and window entry add 14 and 6 bytes. Each place uses a 13-byte header and eight 8-byte intervals, totaling 77 bytes. Together with the 16-byte stream header, these give Equation 1. Table 6 summarizes the packed allocation, and Table 7 gives its field ranges.

The fp16 visual codebook adds $K \cdot 768 \cdot 2 = 1536K$ bytes and is charged in full to each tiered-state AMB cost curve. Static vocabularies and model parameters are shared metadata. Persistent state, runtime workspace, and perception are accounted separately.

Table 6: Fixed-layout implementation of tiered state compaction. Allocation is determined by (M, K).

Component	Capacity	B/row	Payload
Header	1	16	version, flags, lengths
Active entity	M	433	typed entity statistics
Dormant entity	M	14	ID, first, latest, count
Place visit	K	77	header, eight intervals
Partner window	M	6	ID and end time

Table 7: Executable record widths and ranges. The maximum unsigned ID value is reserved for absence.

Quantity	Width	Range	Meaning
Entity or action ID	2 B	0–65,534	Vocabulary index
Place ID	1 B	0–254	Frozen visual bin
Time	4 B	$0–2^{32}-1$	Decisecond ticks
Count or weight	4 B	$0–2^{32}-1$	Integer multiplicity
Action row	14 B	16 rows	ID, first, latest, count
Temporal row	12 B	12 rows	Start, end, weight
Place-count row	5 B	4 rows	Place and count
Partner-count row	6 B	4 rows	Entity and weight

A.4 Record capacity and the storage crossover

Record matching compares retained units, whereas byte matching compares persistent allocations. At 64 records, AMB event tuples use 832 bytes, tiered state uses 29624 bytes, and keyed k-center uses 99,136 bytes. Including its codebook, tiered state costs 41912 bytes, or 40.9 KiB.

Table 8 separates mean recoverable checkpoints from constant reference allocations. The duration study uses the 28.3 KiB tiered ALFRED reference with a fixed layout.

Table 8: Configuration and storage map. Core caps and checkpoint sizes. Tiered rows use a fixed layout and exclude the place codebook.

Configuration	Allocation	State KiB	Evaluation
μWM, fixed precision	16 KiB core	13.38	Count fidelity 85.89%
μWM, fixed precision	24 KiB core	17.58	Count fidelity 95.43%
μWM, fixed precision	48 KiB core	22.30	Count fidelity 100.00%
μWM, adaptive	16 KiB core	13.38	Fourteen-family QA 86.90%
Tiered ALFRED	M=64, K=0	28.33	Scalar and retrieval studies
Tiered AMB	M=64, K=8	28.93	Field, identity, and runtime studies

The 24 and 48 KiB-core variants reach maximum checkpoints of 24.48 and 45.30 KiB. Checkpoint means summarize serialized storage. The largest variant retains all evaluated histories in the shared lossless-prefix stage.

AMB logs average 2.6 KiB and reach 9.1 KiB. The fixed state becomes smaller than a 13-byte tuple log at 2279 events, or 3225 including the codebook. The ALFRED duration sweep crosses its corresponding 2,417-event allowance under K=0, 12-byte tuples, and a 29008-byte state budget.

B Query algebra and formal guarantees

B.1 Fixed memory and scalar reads

Proof of Proposition 1. The empty state satisfies all capacities. Each update restores the bounds on entity tables and nested maps. Keeping at most $M-1$ window entries before appending one

also preserves $|W| \leq M$, and the fixed place vocabulary gives $|V| \leq K$. Fixed-width serialization therefore has the constant allocation in Equation 1, with the byte ledger given in Appendix A.3.

For scalar reads, the online accumulator and full scan start from the same empty value and apply the same update to each matching event. Induction gives equal values after every prefix, assuming required fields remain resident and do not overflow. Reading or composing those fields preserves the equality at the declared timestamp precision.

Table 9: Query compilation. Scalar families fold into fixed accumulators. Interval families use explicitly capped summaries.

Type	Fold	State read and guarantee
Q1	$\min / \max$	entity/action endpoints, exact under residency
Q2/Q3	$\min / \max$	nearest stored endpoint around the query time
Q4	timestamped $\max$	latest place in the resident entity row
Q5	addition	capped partner histogram
Q6	addition	capped place histogram
Q7	weighted merge	12-row temporal coreset
Q8	interval union	eight latest place visits

Reader conditions. Residency includes action, place, and partner subrows, not just the entity row. Partner counts use the same bounded-window pair construction in both computations. Table 5 defines which fields restart after reactivation. Q2/Q3 read stored endpoints, Q7 uses the temporal coreset, and Q8 is exact when all required visits remain in its eight interval rows. Missing keys produce an unanswered query and count as errors.

B.2 Counting and temporal information requirements

Proof of Proposition 2. Consider streams of 0 through H copies of one identical tuple. An event-only buffer can output 0 through M copies, giving $M + 1$ states for $H + 1$ different counts. Exact answers therefore require $M \geq H$. A counter represents the same answers with $\lceil \log_2(H+1) \rceil$ bits. For a zero-error randomized selector, fix an outcome that answers this finite family correctly and apply the same argument. Selectors carrying auxiliary counters fall outside the event-only premise.

Proof of Proposition 3. There are 2^H occupied-bin patterns for one entity. Any two patterns differ at some bin, whose singleton-interval query has different answers. Exact interval queries must therefore distinguish all 2^H states, requiring at least H bits.

Mass conservation and ordering. Each arrival adds unit weight, and merging adds weights, so event mass is conserved within an active epoch. Merged endpoints are convex combinations before rounding. Midpoint order is preserved when input midpoints are nondecreasing. For general overlapping intervals, the updater uses arrival adjacency and the smallest forward midpoint gap. Bounded row count and integer mass conservation hold in either case.

C Evaluation protocol and baselines

C.1 Data interfaces and frozen choices

Each ECMP arm ingests the same entity, action, interval, and visual-place assignments in the same order. The full raw-event index uses the corresponding deterministic readers, isolating the effect of replacing complete history with bounded state or selected events.

AMB provides eight families with different information needs. The five base families are also available on the complete 100-video benchmark. The spatial and visit analyses use the 37-video subset with the required annotations and cached visual features. Six videos are used for design and 31 for confirmation. These roles are separated at the video level because questions from one recording share events, entities, and visual context.

The design videos are P04_05, P06_101, P10_04, P25_107, P28_103, and P30_05. The complete confirmatory list is

P01_103, P02_101, P02_102, P02_121, P02_122, P02_124, P02_128, P02_130
P02_135, P03_101, P03_112, P03_23, P04_101, P04_109, P04_110, P04_114
P04_121, P06_107, P08_16, P11_102, P11_103, P11_16, P18_03, P18_06
P18_07, P22_117, P28_101, P28_113, P28_14, P30_101, P30_111

Mechanism design used six participants—P04, P06, P10, P25, P28, and P30—whose recordings never enter the evaluation cohorts. The precision cohort composes the 96 recordings of six design-disjoint participants—P01 (16 recordings), P02 (23), P03 (20), P08 (10), P11 (4), and P22 (23)—in a fixed order with 120-second gaps. Its fourteen families add action-, place-, and jointly conditioned prefix counts and activity times to the eight base queries. The 672-question sample is fixed without inspecting method outcomes. The full catalog includes 103,662 queries with zeros, of which 66,107 have nonzero counts. The five-family AMB extension uses 16115 questions from 100 videos.

The 96 cohort recordings are

P01 (16): P01_01, P01_02, P01_03, P01_05, P01_06, P01_09, P01_102, P01_104, P01_105, P01_106, P01_107, P01_108, P01_16, P01_17, P01_18, P01_19
P02 (23): P02_01, P02_02, P02_03, P02_04, P02_06, P02_07, P02_103, P02_104, P02_105, P02_107, P02_108, P02_109, P02_110, P02_112, P02_113, P02_115, P02_118, P02_120, P02_123, P02_126, P02_129, P02_131, P02_132
P03 (20): P03_02, P03_04, P03_05, P03_07, P03_09, P03_10, P03_102, P03_107, P03_109, P03_113, P03_117, P03_118, P03_119, P03_120, P03_121, P03_123, P03_14, P03_19, P03_20, P03_28
P08 (10): P08_01, P08_04, P08_05, P08_06, P08_18, P08_21, P08_23, P08_24, P08_26, P08_27
P11 (4): P11_101, P11_104, P11_105, P11_109
P22 (23): P22_05, P22_06, P22_08, P22_09, P22_10, P22_101, P22_104, P22_106, P22_107, P22_108, P22_109, P22_11, P22_111, P22_112, P22_113, P22_115, P22_116, P22_12, P22_13, P22_14, P22_15, P22_16, P22_17

The place model uses cached 1-fps DINOv3 features (Siméoni et al., 2025) and eight fp16 centroids fitted from at most 2,000 normalized features per design video, with seed 0 and eight k-means initializations. The codebook is frozen before confirmatory evaluation. The frozen metadata records NumPy 2.2.6, scikit-learn 1.9.0, 12,288 codebook bytes, and assignments for 47 videos. The 31 confirmatory videos form the evaluation cohort. Figure 10 shows example scenes from the six design videos listed above.

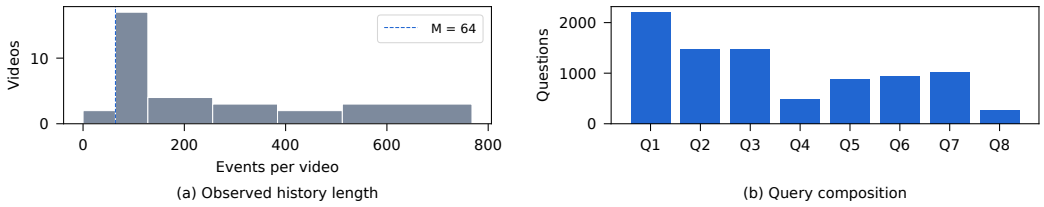


Figure 9: History and query workload. Left, event counts across the confirmatory videos. Right, all 8,729 questions by family.

C.2 Controls, pairing, and uncertainty

For the precision study in Table 3, paired 95% participant intervals for conditional-count gains at 4, 8, and 16 KiB are $[+1.83, +6.53]$, $[+10.03, +23.77]$, and $[+10.07, +31.80]$ percentage points, respectively.

For the tiered-state AMB studies, the estimand is question-weighted paired accuracy. For video v, let N_v be its number of questions and D_v the number correct only under the state minus the number correct only under the event memory. Then

$$\widehat{\Delta}_{\text{question}} = \frac{\sum_v D_v}{\sum_v N_v}.$$

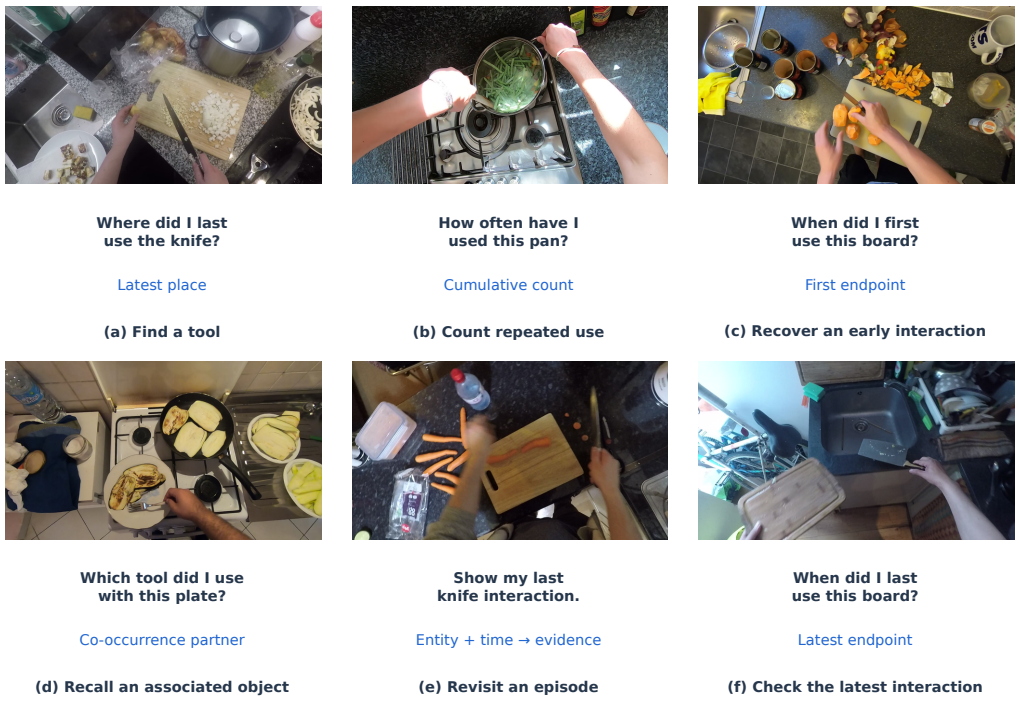


Figure 10: Memory questions across six real kitchens. EPIC-KITCHENS frames from the six design videos illustrate the AMB inputs (Damen et al., 2022; Goletto et al., 2024).

A video-clustered bootstrap samples videos with replacement and keeps every question in a sampled video together. Repeatedly sampled videos contribute repeated copies of their counts. We form 10,000 such samples with seed 0 and report the 2.5th and 97.5th percentiles.

Resampling the same videos in both arms preserves pairing and within-video dependence. The equal-video estimand instead gives each recording equal weight,

$$\widehat{\Delta}_{\text{video}} = \frac{1}{V} \sum_{v} \frac{D_v}{N_v}.$$

The leave-one-video-out gain removes each video while retaining question weighting on the remainder. Both reuse frozen predictions.

C.3 Shared-model memory comparison

The comparison in Table 2 uses native saved memories from SeCom, A-Mem, and HippoRAG 2. Their builders ingest 32-event blocks through frozen local model interfaces. The shared readout uses Qwen2.5-VL-7B-Instruct in text-only mode, greedy decoding, identical statistical tools, and a 64-token answer limit. Event tuples and aggregate records use their corresponding codecs. Original event text is parsed directly. Generated notes are decoded into entity, action, place, time, and count fields without access to evaluation questions. Format retries split note batches without changing their content. Missing facts and unresolved parse failures remain part of the scored result.

Full-store access removes retrieval truncation from this comparison. The storage axis excludes query-time normalization and workspace. Released artifacts contain both tool outputs and final model answers for all 78 configurations on the same 672 questions. Table 2 presents the 13 main configurations.

Middle-budget text selection. Text FIFO and systematic uniform sampling retain complete 32-event blocks in the same textual format used by the recent memory systems. Table 2 shows FIFO at 64, 256, and 512 KiB and uniform sampling at 128 KiB to cover distinct storage regimes. At each cap, every selector uses the largest common block count whose worst-case payload and metadata fit. Reported bytes measure actual occupied text and framing, without padding. Uniform sampling uses a seeded random phase and expands retained counts by the population-to-sample ratio before rounding once to the nearest integer. The ratio is recovered from paid metadata. Construction permits full-corpus access for uniform selection. Query-time tools read all retained blocks from the checkpoint through the shared fourteen-family interface. These rows measure text-memory selection, while the original typed-cache rows measure compact structured encodings of retention policies. The fixed seed, selected blocks, complete checkpoints, and raw model replies are available in `same_tools_midbudget_v1/`.

Figure 1 uses the same stored evidence and prefix-count reader on the full 66,107-question nonzero-count catalog. Its metric is direct integer-count accuracy. Its 26 plotted configurations are drawn from the 121-configuration comparison in `publication_count_candidate_v1/expanded_results.json`, under `artifacts/analyses/agent_memory_baselines/`. The plotted methods comprise fixed-precision μWM, the four-cell joint table, recency and reservoir caches with and without auxiliary statistics, SeCom, A-Mem, HippoRAG 2, VarOpt, and double-precision direct-index t-digest (Cohen et al., 2009; Dunning & Ertl, 2019). The file retains all measured configurations, while the paper discusses the displayed comparisons.

C.4 Pixel-to-answer reference systems

The separate full-AMB reference set contains five accepted 2024–2025 systems. LLoVi is an EMNLP 2024 caption-based framework. AMEGO-L and memory-augmented VideoAgent appeared at ECCV 2024. VideoTree and VideoMindPalace appeared at CVPR 2025 (Zhang et al., 2024; Goletto et al., 2024; Fan et al., 2024; Wang et al., 2025; Huang et al., 2025). Table 10 reproduces the full AMB breakdown from Table 3 of Huang et al. (2025), with the AMEGO-L row independently matched to Goletto et al. (2024). These 100-video, eight-family pixel-to-answer results provide context for ECMP, whose paired comparisons hold semantic events and readers fixed. Table 11 lists the venue, representation, and input of each reference system beside the tiered-state scores under predicted and given identities.

Table 10: Full-AMB reference systems. Percentage scores from Table 3 of Huang et al. (2025), with AMEGO-L also matched to Goletto et al. (2024).

Method	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
LLoVi	22.3	21.4	21.8	22.2	25.6	26.7	18.1	22.2	22.4
AMEGO-L	33.7	36.3	37.2	38.3	27.6	44.3	34.7	48.9	36.3
VideoTree	17.5	25.3	25.0	32.7	34.3	42.5	15.4	22.9	27.0
VideoAgent	35.6	26.1	23.3	34.2	38.2	39.2	38.2	43.5	34.6
VideoMindPalace	**44.7**	27.9	24.2	34.5	**42.7**	**46.4**	**41.0**	**57.4**	**37.8**

Table 11: AMB systems and ECMP reference results. Pixel-to-answer scores follow Huang et al. (2025). ECMP rows use shared event inputs.

Method	Venue	Representation	Input	Videos/types	Acc.
LLoVi	EMNLP'24	Caption summaries	Pixels	100/8	0.224
AMEGO-L	ECCV'24	Tracklets + places	Pixels	100/8	0.363
VideoTree	CVPR'25	Adaptive hierarchy	Pixels	100/8	0.270
VideoAgent	ECCV'24	Temporal + object memory	Pixels	100/8	0.346
VideoMindPalace	CVPR'25	Semantic graph	Pixels	100/8	**0.378**
μWM	Ours	Tiered typed state	Predicted IDs	31/8	0.302
μWM	Ours	Tiered typed state	Given IDs	31/8	0.509
μWM	Ours	Tiered typed state	Given IDs	100/5	0.569

D COMPLETE AMB RESULTS AND ROBUSTNESS

D.1 THE COMPLETE CAPACITY FRONTIER

Table 12 summarizes both ECMP protocols at $M{=}64$. Each protocol uses identical semantic events and readers for the state, the strongest finalized selection control, and the complete raw log.

Table 12: Controlled ECMP comparison. Five-way accuracy and paired 95% video-clustered gain intervals against the strongest event selector in each protocol.

Protocol	Types	Videos	Questions	μWM	Selection control	Raw log	Paired gain [95% CI]
Confirmatory ECMP	8	31	8729	**0.509**	0.405 (k-center)	0.514	**+0.104** $[+0.073, +0.130]$
Complete AMB	5	100	16115	**0.569**	0.430 (entity-strat.)	0.590	**+0.139** $[+0.112, +0.167]$

Table 14 reports all eight capacities, and Figure 11 includes all six selection controls. The state plateaus at 64 slots, while uniform and k-center approach the raw-log score at 512 records.

At $M{=}16$, state coverage is 0.862 against 0.686 for k-center, and accuracy conditional on answering is 0.462 against 0.296. Typed state improves both access to query keys and their retained information.

D.2 QUERY-FAMILY TRANSFER TO THE COMPLETE BENCHMARK

Table 13 gives all eight families. Repeated-use timing, place frequency, and object ordering show the largest gains over the 64-event k-center control.

Table 13: All eight AMB types at $M{=}64$.

Type	Question	N	μWM	k-center	Raw log
Q1	object order	2206	0.594	0.475	0.591
Q2	after anchor	1471	0.524	0.448	0.526
Q3	before anchor	1471	0.494	0.421	0.513
Q4	where left	482	0.274	0.207	0.274
Q5	used with	882	0.305	0.223	0.321
Q6	where used	931	0.377	0.237	0.376
Q7	when used	1012	0.766	0.569	0.770
Q8	when visited	274	0.394	0.401	0.394

The separate 100-video evaluation uses all 16,115 official questions from the five base families. Figure 12 includes every available event control and the raw-log reference. The ordering holds across all five families, extending coverage beyond the 31-video confirmation set. Because the complete benchmark re-includes the six design videos, we also report the design-excluded subset: over 94 videos and 14943 questions, μWM reaches 0.568, the strongest event control reaches 0.434, and the raw log reaches 0.586, with a paired gain of $+0.134$ $[+0.104, +0.162]$.

Table 14: Record-capacity sweep across state slots and retained events.

	μWM		k-center		Event-tuple accuracy				
M	Acc.	KiB	Acc.	KiB	uniform	entity	FIFO	KiB	state–k-center CI
4	0.130	14.4	0.097	6.0	0.096	0.068	0.071	0.05	[+0.024,+0.045]
8	0.243	16.2	0.135	12.1	0.139	0.083	0.086	0.10	[+0.080,+0.142]
16	0.398	19.7	0.203	24.2	0.204	0.134	0.116	0.20	[+0.154,+0.241]
32	0.488	26.8	0.303	48.4	0.311	0.251	0.168	0.41	[+0.164,+0.205]
64	0.509	40.9	0.405	96.8	0.399	0.403	0.265	0.81	[+0.073,+0.130]
128	0.509	69.2	0.473	193.6	0.462	0.456	0.373	1.62	[+0.013,+0.057]
256	0.509	125.9	0.500	387.2	0.499	0.457	0.437	3.25	[-0.001,+0.020]
512	0.509	239.1	0.508	774.5	0.508	0.457	0.487	6.50	[-0.001,+0.004]
Raw log	0.514	–	–	–	–	–	–	–	–

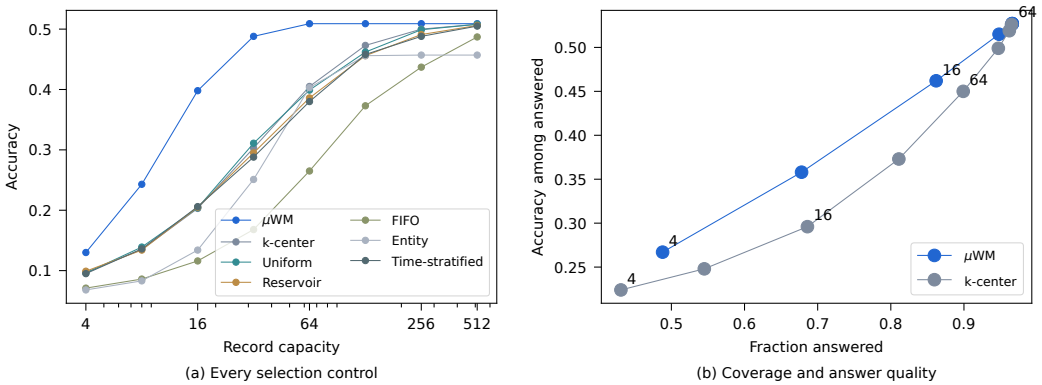


Figure 11: Capacity and answer density. Coverage times conditional accuracy gives overall accuracy.

D.3 Video weighting and leave-one-out sensitivity

The state gains on 26 videos, ties on three, and loses on two. All 31 leave-one-out gains remain positive, ranging from 0.0926 to 0.1074 around the pooled gain of 0.1045. Equal-video weighting gives 0.0760. The video-clustered bootstrap intervals are $[0.0733, 0.1305]$ question-weighted and $[0.0528, 0.1002]$ equal-video. Figure 13 plots both complementary checks.

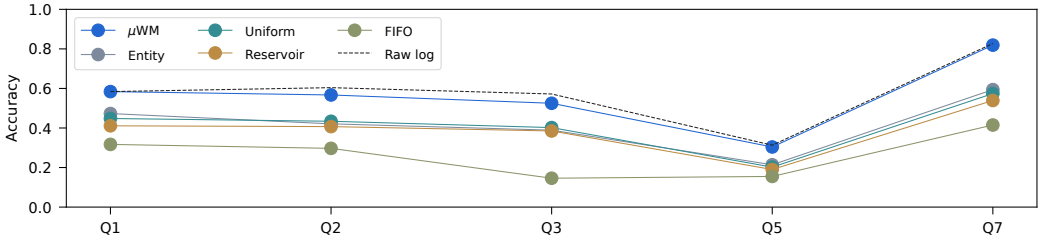


Figure 12: Transfer across 100 AMB videos. Five-family accuracy at $M=64$, with shared semantic events and an unbounded raw-log reference.

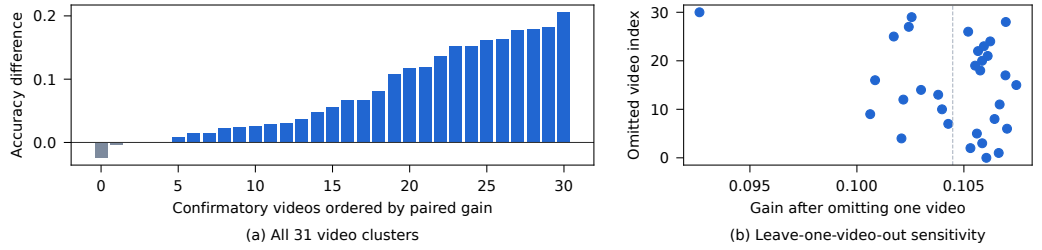


Figure 13: Video-level robustness. Left, all 31 paired video gains. Right, pooled gains after each video omission. The dashed line marks the full-cohort gain.

D.4 Answer preservation after compilation

State and raw-log predictions are joined by video, family, and question ID. The analysis checks unique keys, identical gold answers, and complete coverage of all 8,729 questions. Figure 14 shows the outcome proportions, while Table 15 gives the corresponding counts.

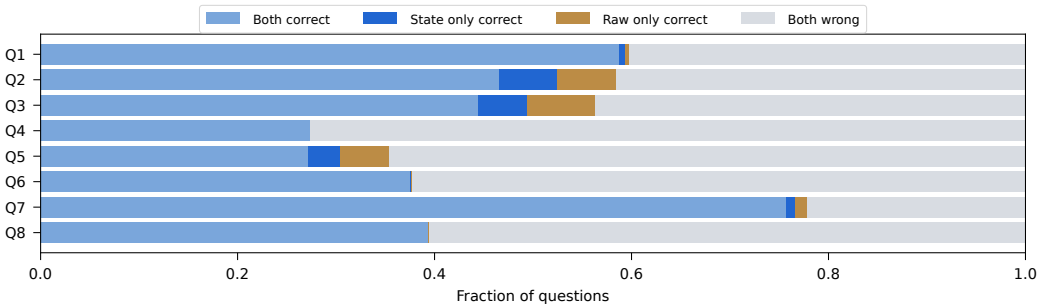


Figure 14: Answer preservation after compilation. Four paired outcomes compare μWM with the complete-event reader across all eight AMB families.

Table 15: Paired answer preservation. Complete outcomes for given identities, frozen ECMP readers, and μWM at M=64, K=8.

Type	Questions	Both correct	Raw only	μWM only	Both wrong
Q1	2,206	1,296	8	14	888
Q2	1,471	685	89	86	611
Q3	1,471	653	101	74	643
Q4	482	132	0	0	350
Q5	882	240	43	29	570
Q6	931	350	0	1	580
Q7	1,012	766	13	9	224
Q8	274	108	0	0	166
All	8,729	4,230	254	213	4,032

The state preserves 4,230 of the raw log's 4,484 correct answers, giving 94.3% question-level retention, and answers 213 additional questions correctly. Q4 and Q8 preserve every reference success. The complete paired counts in Table 15 distinguish answer retention from the 99.1% ratio of aggregate accuracies.

D.5 Recognition and design-only prior selection

Both memories receive identical predicted noun IDs, with boundaries, actions, places, queries, and readers unchanged.

The recognizer combines zero-shot SigLIP-large, zero-shot SigLIP2-so400m, and a fine-tuned tower on contacted-object crops. Scores are standardized per event. The fixed fusion weights are 0.5 for the fine-tuned tower, 1.5 for SigLIP-large, and 1.0 for SigLIP2. A noun-given-action prior is fitted on training videos disjoint from all 37 evaluated videos.

The prior weight is selected from $\{0, 0.5, 1, 1.5, 2, 3\}$ by memory QA accuracy on six design videos. Weights 1 and 1.5 tie at 469/1,550 correct, so the smaller-weight rule freezes 1 (Figure 15).

Predictions use the same ontology mapping as annotated IDs. Released video/time/class records replay the intervention without the visual towers.

D.6 Paired gains under the noisy channel

The frozen recognizer reaches 0.340 event-level noun accuracy on the confirmatory data. With that shared channel, the state reaches 0.302 and the matched event memory reaches 0.246. The paired difference is 0.056 with a video-clustered interval $[0.041, 0.070]$ over all 8,729 questions (Table 16).

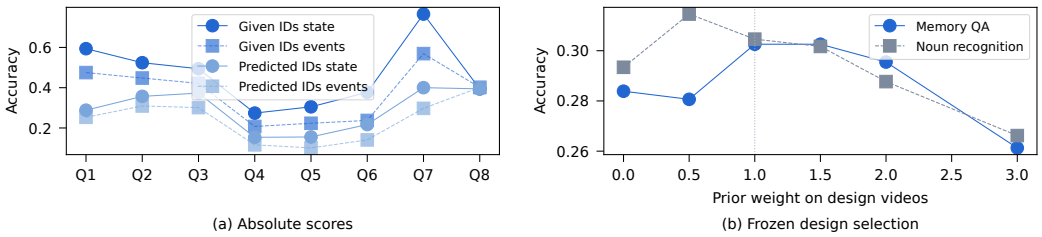


Figure 15: Predicted-identity transfer. Left, state and event-memory scores under both noun channels. Right, the design-set prior sweep with selected weight 1.

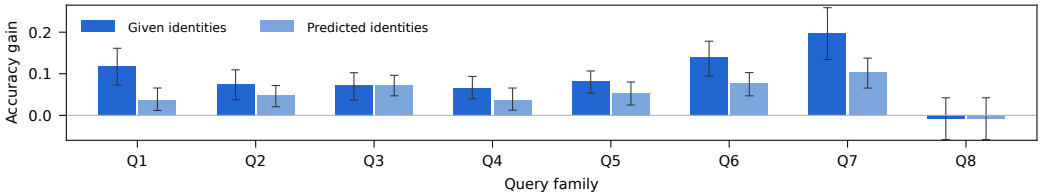


Figure 16: Query-family gains under both identity channels. Paired accuracy differences at $M{=}64$, with 95% video-clustered bootstrap intervals.

Table 16: Identity-channel transfer. Both channels use the same 31 videos, eight query families, reader rules, and 64-event k-center control.

Noun channel	Noun acc.	μWM	k-center	Gain [95% CI]
Given semantic IDs	1.000	0.509	0.405	$+0.104$ $[+0.073, +0.130]$
Video-disjoint predictions	0.340	0.302	0.246	$+0.056$ $[+0.041, +0.070]$

Figure 16 shows positive gains on Q1–Q7. Across all questions, 821 are correct only under state and 328 only under the event memory, versus 1,444 and 532 with given identities. Q8 uses place visits, so noun substitution leaves it unchanged. The gains on Q1–Q7 show that the same typed state remains useful with a predicted identity channel.

E Field mechanisms and temporal capacity

E.1 Field contributions and query dependencies

Each ablation removes one component while keeping input and reader rules fixed. Table 17 reports overall effects, and Figure 7(a) gives the per-family changes.

Table 17: ECMP field ablations. $M{=}64$ isolates query fields. $M{=}16$ exposes dormant-tier behavior under entity pressure.

M	State	Acc.	Δ	Primary family
64	Full μWM state	0.509	–	all eight
64	remove Location fields + visit table	0.441	-0.068	Q4/Q6/Q8
64	remove 12-point temporal coreset	0.482	-0.027	Q7
64	remove Per-verb first/last/count	0.484	-0.025	Q1/Q2/Q3
64	remove Partner map	0.506	-0.003	Q5
64	remove dormant summary	0.507	-0.002	long-tail entities
16	Full μWM state	0.398	–	all eight
16	remove dormant summary	0.283	-0.115	capacity pressure

Location removal affects Q4/Q6/Q8, temporal removal lowers Q7 by 0.231, and partner removal lowers Q5 by 0.034. Action rows primarily support Q1–Q3, matching the reader dependencies.

Q5 falls back to action-interval proximity when its partner map provides no positive count, explaining its small response to removing action rows. The fallback is fixed across conditions.

E.2 Dormant-tier history coverage

Dormant storage matters under entity pressure. The five-capacity replay uses the same 31 videos, 8,729 questions, and frozen eight-place codebook. At each capacity, we remove the dormant summary before readout while retaining the compiled active state and reader rules. The accuracy gains are 0.087, 0.115, 0.048, 0.014, and 0.002 at $M{=}8, 16, 32, 48, 64$, respectively. The largest absolute gain occurs at $M{=}16$, then falls as active capacity increases. Existing full-state predictions and both published removal conditions match the replay question by question. The paired predictions are in `artifacts/analyses/dormant_capacity_v1/`. Appendix G.2 isolates this effect by cycling through working sets larger than the active table.

E.3 Temporal design and disjoint confirmation

The temporal study varies interval capacity while keeping a summary for every observed entity. It therefore isolates temporal compression from entity eviction. The design set contains 214 Q7 questions from the six design videos. Capacities are 2, 4, 8, 12, and 16. Each capacity compares weighted merging, the most recent intervals, and reservoir sampling over five fixed seeds.

At twelve rows, the merge scores 0.874 and the reservoir mean is 0.875. Sixteen merge rows reach the full-interval reference of 0.879. Choosing twelve saves $4 \cdot 12 = 48$ bytes per active entity, or 3,072 bytes across 64 slots, for a small observed design-set difference. This choice is frozen before confirmation.

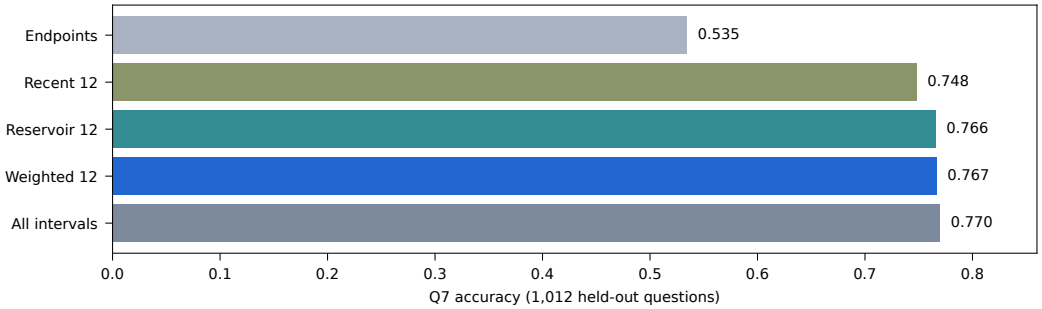


Figure 17: Compact temporal recall. Twelve-row summaries evaluated on 1,012 held-out Q7 questions. Reservoir uses frozen seed 0.

On the confirmatory videos, twelve merged intervals score 0.7668, reservoir scores 0.7658, and recent intervals score 0.7480. The unbounded reference scores 0.7698. The endpoint-only reader scores 0.5346. Compact temporal support supplies the main gain. Merging and reservoir are close at this capacity, while merging conserves mass deterministically. The integrated state scores 0.766 after applying entity residency and packed timestamp conventions.

F Household transfer and long-history scaling

F.1 Transfer on natural household trajectories

ALFRED trajectories become ordered noun/action intervals (Shridhar et al., 2020). Five generated families ask about first-of-four object order, relative first occurrence, exact counts, latest use, and the latest object. The untiled study contains 105 scenes and 4,175 questions, with 414 mean and

1,302 maximum events. With no visual-place queries, the state uses $K{=}0$. At 64 slots it preserves all generated answers, while the strongest 64-event selector reaches 0.608 (Table 18).

Table 18: Untiled ALFRED household trajectories. Accuracy over five statistic-query types. The final row retains every event.

M	μWM	uniform	entity	tracklet	FIFO
16	0.840	0.280	0.323	0.327	0.320
64	1.000	0.415	0.608	0.481	0.470
256	1.000	0.815	0.608	0.853	0.832
∞	1.000	–	–	–	1.000

F.2 Equal-byte scaling across every selector

The duration study repeats each complete source stream by factors $1, 2, 4, 8, 16, 32$, adding a 20-second gap between copies. It preserves vocabulary and within-tile event order. Questions are regenerated after tiling, so a count answer includes every copy and an endpoint answer refers to the resulting full history. The artifact records the generated question count at each factor.

The tiered reference keeps a fixed layout at $M{=}64, K{=}0$, giving 29008 bytes. Four tuple selectors receive 2,417 events. The tracklet reference receives 2,417 complete groups and can store more events. Figure 18 reports all five controls.

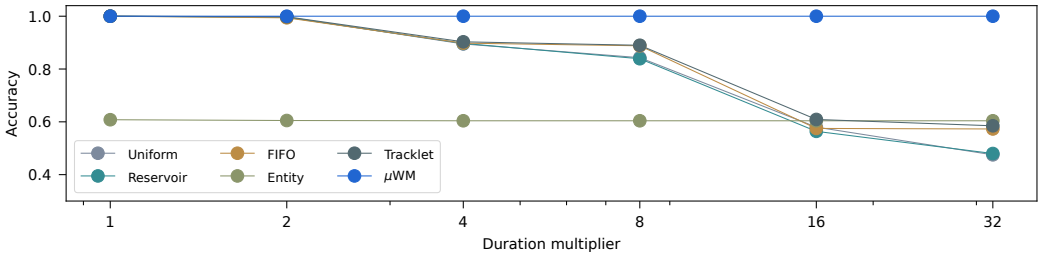


Figure 18: Retention across duration. The same 105 source scenes compare tiered-state μWM with five selectors at every replay factor.

Table 19: Fixed-state duration scaling. Event-control best is the strongest of five selectors at each factor. Gain intervals cluster the 105 source scenes.

Tiling	Mean events	Raw mean KiB	μWM	Event-control best	Difference [95% CI]
1	414.0	4.9	1.000	1.000	[+0.000,+0.000]
2	828.1	9.7	1.000	0.998	[+0.000,+0.007]
4	1656.2	19.4	1.000	0.903	[+0.066,+0.129]
8	3312.3	38.8	1.000	0.890	[+0.078,+0.145]
16	6624.6	77.6	1.000	0.609	[+0.366,+0.414]
32	13249.2	155.3	1.000	0.604	[+0.382,+0.411]

The mean history exceeds the 2,417-event allowance at $8\times$. At $32\times$, the strongest control scores 0.604 while the state remains exact. The paired gain interval is $[0.382, 0.411]$, clustered by the 105 source scenes. Exact raw logs grow to 155.3 KiB on average and 488.2 KiB at maximum.

F.3 Early, cumulative, and recent-history retention

Figure 19 separates early, cumulative, and recent history. Independent first, latest, and count fields preserve these statistics without competing for positions in an event buffer.

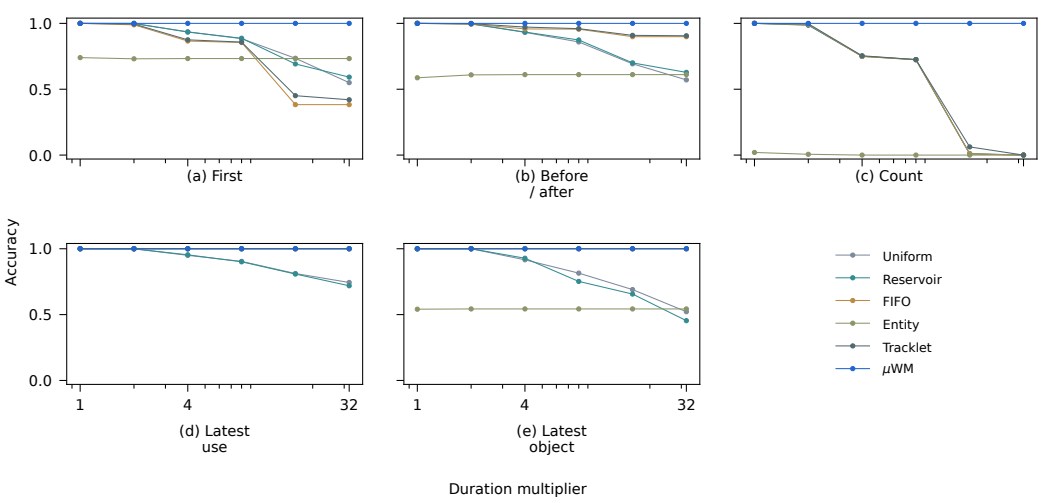


Figure 19: Query-family retention across duration. Early, cumulative, and recent-history accuracy for μWM, four tuple selectors, and a tracklet reference.

At $32\times$, all five tested subsets score zero on exact counts. FIFO and tracklet selection still answer latest-use and latest-object questions perfectly, yet score 0.383 and 0.420 on first-of-four ordering. Entity stratification raises first-of-four accuracy to 0.733 but reaches 0.543 on latest-object questions.

F.4 State-directed routing for description retrieval

The pilot uses all 251 seen and 255 unseen ALFRED validation trajectories across 88 seen and four unseen environments. Its protocol is fixed before either split. Episodes concatenate scene-locally with seed 7, five-second action steps, and twenty-second gaps. The first human annotation describes each high-level action. The synthetic timestamps encode action order. Figure 20 illustrates the archive's scene diversity.

Queries cover every repeated object/action pair with an annotated latest event. Composed programs first select an action's unique most-frequent object, excluding frequency ties. This yields 241 queries in 53 seen environments and 118 in four unseen environments. One undescribed action is ineligible as a target but still enters state. Structured query programs supply the entity and action keys.

Table 20: State-directed evidence routing. Hit@k measures source-description recovery. Work counts mean descriptions scored per query.

Retriever interface	Hit@1	Hit@3	Work
Seen environments, 241 queries			
Full lexical	0.270	0.593	33.0
State rewrite only	0.336	0.656	33.0
μWM gate	0.813	0.992	4.1
Full-history scan gate	0.813	0.992	4.1
Unseen environments, 118 queries			
Full lexical	0.076	0.144	694.6
State rewrite only	0.110	0.186	694.6
μWM gate	0.644	0.958	4.1
Full-history scan gate	0.644	0.958	4.1

All arms use character TF-IDF with 3–5-character word-boundary features, sublinear term frequency, and newest-first ties. Rewriting resolves action and noun but searches all descriptions. Gating adds a ten-second window around the stored endpoint. The scan control obtains the same keys from full history. Missing keys or empty gates trigger full-archive fallback. All 359 state routes match the scan, with no fallback (Table 20).

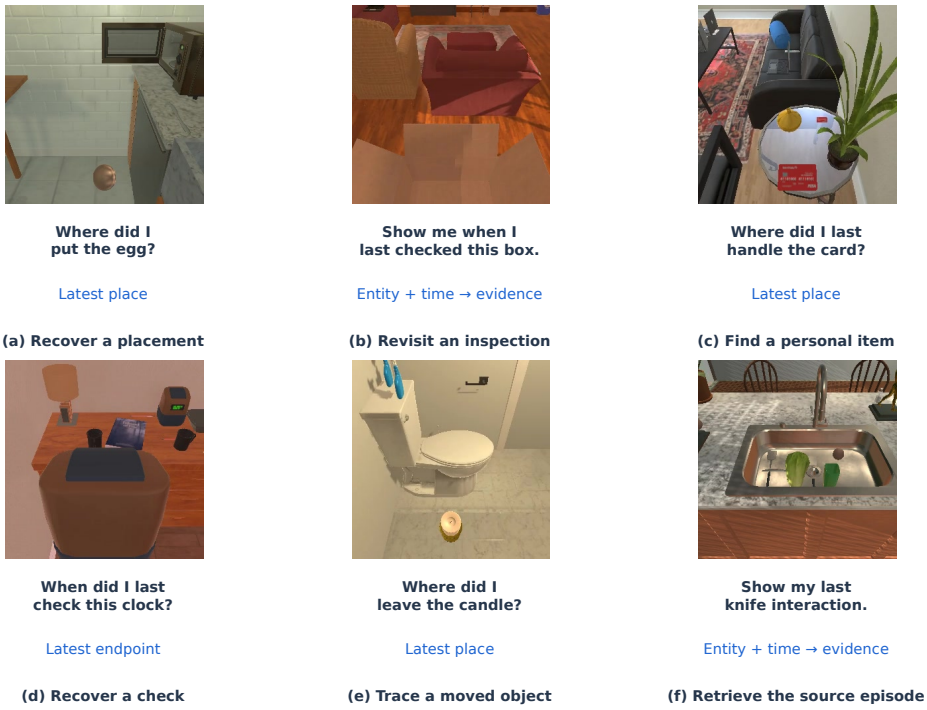


Figure 20: Household evidence across four room types. Native ALFRED frames from six validation trajectories illustrate the description-retrieval archive (Shridhar et al., 2020).

State remains 29008 bytes, with at most 29 nouns per stream. External evidence is separately charged: per-stream description payloads and sparse TF-IDF indices range from under a KiB to 742,488 bytes across splits, excluding interpreter overhead. The archive and its indices grow with history.

The composed queries achieve 0.986 support Hit@3. Every state-derived anchor matches its full-history counterpart, and the retriever returns an extractive description from the gated evidence. Across all queries, 7 target descriptions remain outside the top three after lexical ranking. Thus the compact state preserves the full-history routing decision while focusing retrieval on a small local evidence set.

G System efficiency and reproducibility

G.1 Latency and serialized growth

The profiling unit is one update call or one deterministic answer call. The released benchmark uses three repetitions and a pinned Intel Xeon Platinum 8470Q CPU core with `perf_counter_ns`. The update distribution contains 8,373 measurements and the query distribution contains 4,650 measurements. Update mean, median, p95, and p99 are 10.45, 10.12, 17.04, and 20.89 microseconds. Query median, p95, and p99 are 29.20, 170.76, and 481.58 microseconds.

Figure 21 plots latency and serialized growth. State stays at 29624 bytes through 100000 events, while bare 13-byte tuples reach 1,300,000 bytes. The separate visual-key curve includes the selection features.

With nested capacities fixed, updates take $O(M)$ work and queries depend on the family and option count, not the original stream length. The profiling scope is the memory update and read path.

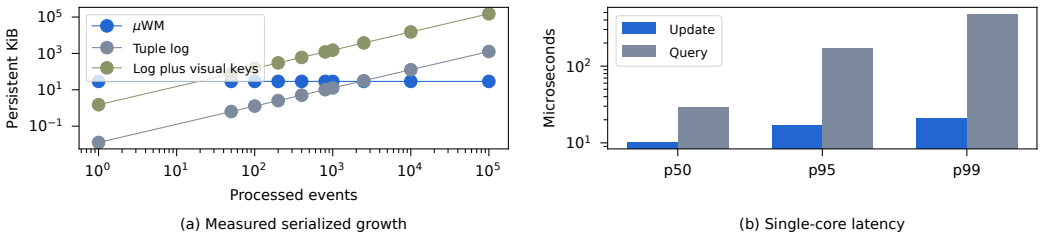


Figure 21: Memory-layer cost. Left, serialized storage through 100,000 events. Right, update and query latency on a pinned Xeon CPU core.

G.2 Resident-history and working-set checks

The released updater cycles through 12 entities, four actions, and four places under $M{=}16, K{=}4$. At 1, 16, 64, 256, 1,024, 4,096, and 16,384 events, independent accumulators check first/latest/count triplets, temporal mass, and serialized size.

Every checked scalar triplet agrees, temporal mass is conserved, and state stays at 7,572 bytes (Figure 22).

A separate working-set sweep cycles ten times through 8, 12, 16, 20, 24, 28, 32, 36, or 40 entities. Every entity should have final count 10. The full state has 16 active and 16 dormant rows. The active-only condition clears the dormant table after each update. The final metric is the fraction of all cycle entities whose retained lifetime count equals 10, with missing entities counted as incorrect.

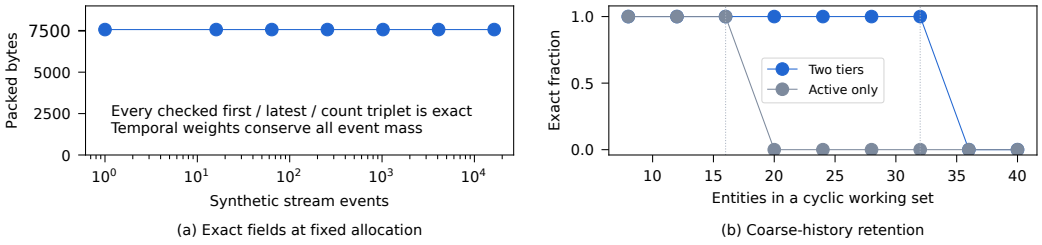


Figure 22: Executable state invariants. Synthetic sweeps test scalar preservation and constant allocation (left), and dormant-tier lifetime counts under working-set growth (right).

Dormant storage doubles exact coarse-history coverage from 16 to 32 entities. The 36- and 40-entity settings extend the sweep beyond the combined slot capacity. Model-free unit tests additionally verify record widths, dense-arrival caps, temporal mass, truncation flags, and overflow rejection.