AMBER: ANYTIME MULTIMODAL MEMORY BRIDGING EPISODES VIA RICH ASSOCIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal conversational memory systems struggle with required evidence when information is **updated, extended, or contradicted** across episodes, where previous similarity-based retrieval failed. To address this challenge, we introduce **AMBER** (Anytime Multimodal Memory Bridging Episodes via Rich Association), which builds graph with temporal, procedural, complementary, and conflicting associations, continually organizing episodes into paths centered on shared association anchors. When association retrieval is needed, AMBER adaptively activates one type for the query, traverses matching paths from initial semantic matches, and ranks the resulting evidence chains under a fixed budget. To evaluate whether memory systems can preserve required evidence while incorporating new information, we introduce **MemStream**, annotated 1,603 information states for 779 evolution-related questions with reference answers and supporting evidence from Mem-Gallery and H2HMem. MemStream enables streaming evaluation with process-level metrics for answer continuity and degradation, revealing an average 58.48% decrement in state F1 compared to the final. Across 2 widely used MM benchmarks and our proposed one, AMBER improves average F1 by 2.68% over the strongest of four baselines, while recovering 19.07% more evidence in cross-episode queries.

1 INTRODUCTION

Imagine using an multimodal large language model (MLLM) assistant to document a baking course, saving photos and messages as you prepare ingredients, knead the dough, and bake the bread (Fig. 1). Weeks later, you ask it to recall the full process so you can bake the bread again. The assistant remembers the introductory, but misses the steps that turn them into a loaf. Without those scattered episodes, you cannot reliably recreate what you learned. This illustrates a central challenge for multimodal conversational memory: connecting required evidence across episodes and retrieving the complete set of details needed to answer a question.

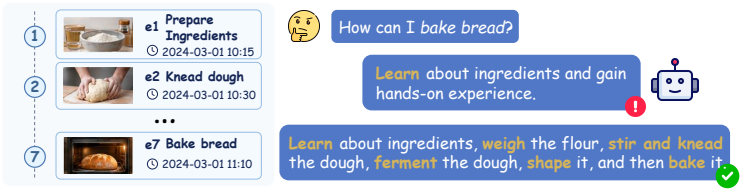

Figure 1: **Examples of required evidence across episodes.** Retrieving only the introductory episode would miss the procedural evidence.

Existing approaches to evidence selection largely fall into two categories. *Similarity-based retrieval* ranks episodes independently against the query (Chen et al., 2022). UniversalRAG additionally routes retrieval by modality and granularity (Yeo et al., 2025), and V-Mem by query and target-evidence modality (Kang et al., 2026). Under fixed budgets, newer similar competitors will obscure earlier required evidence. While evidence relevance may depend on source context (Sun et al., 2026), *graph-based memory* connects records through tags, concepts, or entities (Xu et al., 2025; Jain et al., 2025; Gutiérrez et al., 2025). It finds evidence through connections: AUGUSTUS (Jain et al., 2025) searches relevant tags, HippoRAG 2 (Gutiérrez et al., 2025) uses query-selected seeds, and MRAgent (Ji et al., 2026) conditions traversal and pruning on the query and accumulated evidence.

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.

However, three challenges remain. **First, tags or shared attributes are still insufficient for modeling cross-episode associations.** Current connections can not distinguish cross-episode roles. They can not specify whether episodes precede, enable, extend, or contradict one another. **Second, query-relevant episodes do not specify which traversal should follow.** Expanding across association types can consume the budget with irrelevant episodes. Traversal needs a query-specific association choice. **Third, cross-episode failures require process-level diagnosis.** Our preliminary evaluation reveals that memory systems struggle to connect evidence as information evolves: across four baselines, mean token-level F1 on cross-episode queries under `gpt-4.1-nano` is 16.91 percentage points lower than on Mem-Gallery (Bei et al., 2026) and H2HMem (Zhu et al., 2026) questions overall. Mem-Gallery also reports difficulties in test-time resolution and conflict detection. Deficits can originate in episode representations, association construction, traversal, or ranking, which final scores cannot disentangle. Hippocampal pattern completion shows that a partial cue can reactivate other elements bound within an event (Horner et al., 2015), motivating the use of association anchor as an entry point to associated evidence. Discourse-relation frameworks further distinguish temporal, contingency, comparison, and expansion relations (Prasad et al., 2008). These distinctions inform our association types for organization. In particular, cross-episode evidence is retrievable only when the relevant episodes connect successfully. It is therefore crucial to examine whether earlier required evidence can still be associated with newly arrived episodes, a metric that final-state scores alone cannot capture.

We therefore introduce **AMBER** (Anytime Multimodal Memory Bridging Episodes via Rich Association) and **MemStream** to address these three challenges. **For association construction**, AMBER builds its graph online through incremental updates as episodes arrive, distinguishing four cross-episode association types: temporal, procedural, complementary, and conflicting. An anchor is a textual objective with which a stored path or retrieved evidence chain is associated. **For retrieval**, association-routed retrieval starts from semantic seeds and activates one query-selected association type to traverse matching paths. The resulting evidence chain, retrieved paths centered on an anchor, are ranked under a fixed budget. association-aware evidence Assembly retains their membership and internal order in the evidence context. On the public Mem-Gallery and H2HMem benchmarks, AMBER achieves state-of-the-art performance. **To evaluate** the challenge of retrieving evidence as information evolves, we construct MemStream from Mem-Gallery and H2HMem, distinguishing information states for each question and annotating each reference answer for evaluation with streaming inputs and process-level metrics, and AMBER still outperforms prior methods. Against four retrieval baselines on these three datasets, AMBER's gains in F1, smoothed unigram bilingual evaluation understudy (BLEU-1) (Papineni et al., 2002), exact match (EM), and LLM-as-a-Judge (LLM-J) scores over the strongest baseline per setting and metric average 2.68%, 2.69%, 1.51%, and 1.47%, respectively, with all settings equally weighted. On MemStream's criteria, AMBER raises supporting-evidence recall among up to ten retrieved episodes (Support R@10) from the strongest baseline's 51.41% to 65.46% and recall of still-valid earlier evidence (Retained R@10) from 56.28% to 68.85%. Taken together, these results show that AMBER consistently improves answer quality while making evolving evidence more accessible.

In summary, our contributions are as follows.

- We propose AMBER, a multimodal conversational memory framework that organizes episodes into a graph with explicit cross-episode associations, enabling an MLLM to reason over specific evidence chains.
- We develop MemStream, a benchmark with streaming inputs, state-specific probes, and state-aware metrics for diagnosing failures in connecting evidence across evolving states.
- Extensive experiments show that AMBER improves answer F1 over the strongest baseline by 2.56%, 2.25%, and 3.23% on Mem-Gallery, H2HMem, and MemStream, respectively, with recovering 74 of the 388 pieces of required evidence that similarity retrieval missed.

2 Related Work

Multimodal Conversational Memory. Long-term memory systems store conversation histories as summaries, notes, entities, or retrieved episodes to support personalized interaction (Xu et al., 2022; Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; LangChain AI, 2026; Nan et al., 2025). Related methods use multimodal similarity search (Chen et al., 2022), modality routing (Yeo

et al., 2025; Kang et al., 2026), or structured memory access (Xu et al., 2025; Fisher, 2025; Jain et al., 2025; Nie et al., 2026), limiting them from retrieving evidence connected through cross-episode associations. AMBER instead organizes episodes over the association graph, dynamically activates the associations to retrieve evidence, and thus bridges this gap.

Graph-Based Retrieval. Graph-based retrieval organizes information through entities, topics, temporal order, and semantic relations (Edge et al., 2024; Gutiérrez et al., 2024; Li et al., 2024). Related database research studies keyword search, subgraph matching, and path queries over structured data (He et al., 2007; Chen et al., 2009; Tran et al., 2009; Neumann & Weikum, 2008; Zou et al., 2011; Libkin et al., 2016). Memory systems further employ graph structures for temporal organization and multi-hop retrieval (Rasmussen et al., 2025; Xu et al., 2025; Sun et al., 2026; Wang et al., 2026a). APEX-MEM stores events and temporally qualified facts (Banerjee et al., 2026). MRAgent conditions graph expansion (Ji et al., 2026), while MAGMA combines semantic and entity graphs (Jiang et al., 2026). While they do not jointly distinguish the cross-episode associations, AMBER enriches associations across multiple types, enabling it to handle complex situations beyond a single association.

Information Evolution and Evaluation. Prior work maintains changing information through consolidation, revision, replacement, and temporal validity (Chhikara et al., 2025; Ong et al., 2025; Su et al., 2026). Temporal databases, provenance, and stream processing provide related perspectives on information validity and queries over changing data (Dyreson et al., 1994; Snodgrass, 2012; Buneman et al., 2001; Arasu et al., 2006). Recent benchmarks evaluate memory over long-horizon and multimodal conversations (Maharana et al., 2024; Wu et al., 2024; Bei et al., 2026; Zhu et al., 2026). Evo-Memory evaluates experience reuse under streaming task inputs (Wei et al., 2026), while STALE tests responses to implicit memory invalidation (Chao et al., 2026). MemConflict evaluates conflict handling (Tao et al., 2026). SubtleMemory evaluates memory relations (Wang et al., 2026b). Among them, MemStream first measured cross-episode evolution in the multimodal conversational task, revealing its inability to handle dynamic changes in information.

3 METHOD

3.1 PROBLEM DEFINITION

Given a multimodal conversation history $\mathcal{D} = (e_1, \ldots, e_N)$, where N is the total number of memory episodes, we study question answering as these episodes arrive. Each conversation session yields one or more episodes. Episode $e_i = (x_i, I_i, c_i, \tau_i)$ contains text x_i, an image set I_i, image captions c_i, and a timestamp τ_i; i indexes episodes in arrival order. At query index t, only $n_t \leq N$ episodes have arrived. The available history is therefore the prefix $\mathcal{D}_{1:n_t} = (e_1, \ldots, e_{n_t})$. Query $q_t = (x_t^q, I_t^q)$ consists of question text x_t^q and an optional image set I_t^q.

The objective is to answer q_t using evidence selected from this prefix. Let $\widehat{\mathcal{E}}_t$ denote the set of selected evidence episodes and K the maximum number of episodes supplied to the answer model. Their text, images, captions, and timestamps form the structured multimodal evidence context $\mathcal{X}_t$. An answer model f_{ans} generates the predicted answer $\hat{a}_t$ from the query and this context:

$$\hat{a}_t = f_{\text{ans}}(q_t, \mathcal{X}_t), \qquad \widehat{\mathcal{E}}_t \subseteq \{e_i : 1 \leq i \leq n_t\}, \qquad |\widehat{\mathcal{E}}_t| \leq K. \tag{1}$$

Successful selection must collect currently required evidence within budget K, including earlier evidence that remains necessary as new episodes arrive.

3.2 OVERVIEW

AMBER couples memory construction and retrieval to maintain access to required evidence as new episodes arrive (Fig. 2). **Online Association Construction** incrementally organizes arriving episodes into a memory graph. After processing e_i, this graph contains the observed episodes $\{e_1, \ldots, e_i\}$ and a set of stored paths Π_i. Each stored path records an association type and an *anchor*, a textual description of what its episodes concern. Incremental path updates connect earlier evidence to subsequent steps, supplements, or corrections. **Association-Routed Retrieval** uses the query to select an association type and anchor, then expands semantically retrieved episodes along matching

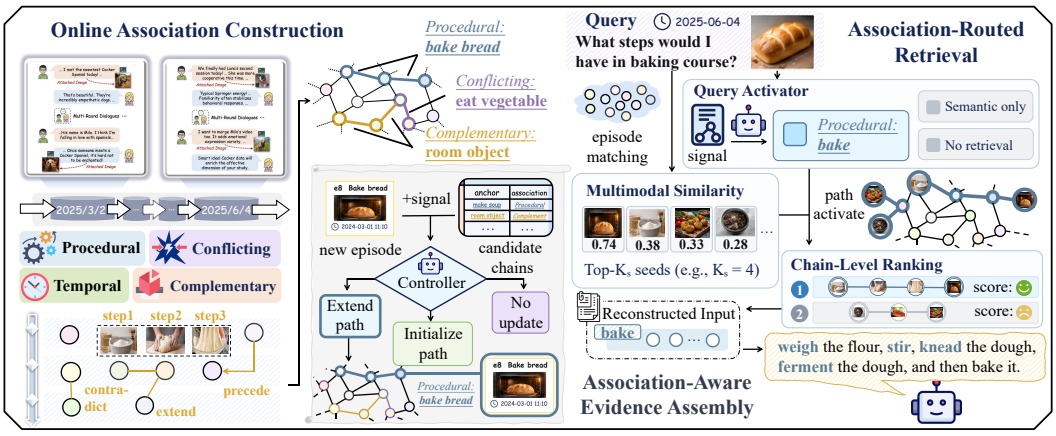

Figure 2: **Overview of AMBER.** Online association construction organizes episodes through explicit four associations. Association-routed retrieval selects which paths to retrieve from semantic seeds. Evidence chains are then ranked, selected, and assembled for answer generation.

paths. The resulting *evidence chains* are query-specific paths centered on an anchor and ranked as units under the retrieval budget. Finally, **Association-Aware Evidence Assembly** forms $\mathcal{X}_t$ from the selected contents, retaining chain order and adding association and anchor descriptors.

3.3 Online Association Construction

A-Mem uses an LLM to link semantically retrieved notes through shared attributes (Xu et al., 2025), while AUGUSTUS connects semantic tags to conversation contexts (Jain et al., 2025). For cross-episode retrieval, topical similarity alone does not distinguish a further step from a correction. AMBER records an explicit association type and anchor during construction, so retrieval can select paths according to both their semantic role and what they concern.

The association inventory $\mathcal{R}$ comprises four roles. A *temporal association* specifies that one episode *precedes* another through explicit ordering or time reference. A *procedural association* specifies that one step *enables* another within the same procedure. A *complementary association extends* or *fills in* information about the same anchor without contradiction. A *conflicting association contradicts* earlier information; a later assertion *supersedes* it when it provides a replacement. These non-exclusive operational types draw on discourse-relation distinctions (Prasad et al., 2008; 2014).

Each stored path has a positive index j and is represented by $\pi_j = (r_j, a_j, \mathcal{V}_j, \mathcal{A}_j)$, where $r_j \in \mathcal{R}$ is its association type, a_j its anchor, $\mathcal{V}_j$ its episode set, and $\mathcal{A}_j$ its directed links encoding path order. For an arriving episode e_i and each type $r \in \mathcal{R}$, AMBER retrieves similar earlier episodes and collects a bounded candidate set $\Pi_i^{r,\mathrm{cand}} \subseteq \Pi_{i-1}$ of their type-r paths. Their indices form $\mathcal{J}_i^r = \{j : \pi_j \in \Pi_i^{r,\mathrm{cand}}\}$.

Let $\sigma^r(x_i)$ measure the strength of fixed signal for type r, and $\boldsymbol{\sigma}_i = (\sigma^r(x_i))_{r\in\mathcal{R}}$ collect these four scores. An LLM controller f_c uses the new text, timestamp, signals, and candidate summaries to produce a validated update decision d_i^r and an anchor a_i^r when initialization is requested:

$$(d_i^r, a_i^r) = f_c(x_i, \tau_i, \boldsymbol{\sigma}_i, \Pi_i^{r,\mathrm{cand}}; r), \qquad d_i^r \in \mathcal{J}_i^r \cup \{0, \perp\}. \tag{2}$$

A positive d_i^r extends that path, 0 requests initialization at anchor a_i^r, and $\perp$ leaves that type unchanged. Each episode belongs to at most one path per type, while independent decisions allow multiple types to coexist.

3.4 Association-Routed Retrieval

MAGMA uses query-dependent weights for graph transitions (Jiang et al., 2026), while MRAgent iteratively expands and prunes memory using the query and accumulated evidence (Ji et al., 2026). In our graph, a seed can belong to paths with different association types and anchors, so expansion

must identify which paths supply the requested evidence. AMBER makes this choice through one selected association type and a generated query anchor, then ranks chains as units to retain associated evidence with weak direct query similarity.

An LLM activator uses query text and the same signal to select a mode $z_t \in \mathcal{R} \cup \{0, \perp\}$ and, for association modes, generate a query anchor a_t^q. Here, 0 requests semantic retrieval alone and $\perp$ indicates that the current input suffices without memory access. When retrieval is required, AMBER forms the seed set $\mathcal{S}_t$ from up to K_s available episodes with the highest multimodal similarity $s_{\mathrm{mm}}(q_t, e)$ between query q_t and episode e. Thus, K_s controls the number of entry points for path expansion, while K limits the final evidence episodes.

Let $s_a(a_t^q, a_j) \in [0, 1]$ measure agreement between query and path anchors, and δ_a be the minimum agreement required for expansion. For an association mode $z_t \in \mathcal{R}$, the activated path set Π_t^{act} contains paths from the current memory Π_{n_t} that include a seed, match the selected type, and pass this anchor threshold:

$$\Pi_t^{\mathrm{act}} = \{\pi_j \in \Pi_{n_t} \mid \mathcal{V}_j \cap \mathcal{S}_t \neq \emptyset, \quad r_j = z_t, \quad s_a(a_t^q, a_j) \geq \delta_a\}. \tag{3}$$

Each activated path π_j becomes a candidate chain $\mathcal{C}_{j,t}$, containing all its episodes in path order and inheriting anchor a_j; the index j identifies its source path. Write $\mathcal{V}(\mathcal{C})$ for the episodes in any candidate chain $\mathcal{C}$. Seeds sharing a path produce one candidate; unmatched seeds remain singletons. With mixing weight $\lambda \in [0, 1]$, the score ρ_t of an associated chain combines its strongest semantic seed and anchor agreement:

$$\rho_t(\mathcal{C}_{j,t}) = \lambda \max_{e \in \mathcal{V}(\mathcal{C}_{j,t}) \cap \mathcal{S}_t} s_{\mathrm{mm}}(q_t, e) + (1 - \lambda) s_a(a_t^q, a_j). \tag{4}$$

An unmatched singleton receives its semantic score alone. When $z_t = 0$, all seeds remain singletons; when $z_t = \perp$, the evidence set is empty. Ranking chains as units allows a relevant seed to bring in associated evidence with weak direct query similarity. AMBER greedily selects chains in descending score order under budget K, retaining internal order and cropping an oversized first chain around its strongest seed. Let $\mathbb{C}_t^*$ be the resulting collection of selected chains, including any cropped chain. It supplies $\widehat{\mathcal{E}}_t = \bigcup_{\mathcal{C} \in \mathbb{C}_t^*} \mathcal{V}(\mathcal{C})$ for evidence assembly. Taxonomy scope and implementation details are provided in Appendices A and D.

4 MemStream Benchmark

4.1 Motivation

MemStream evaluates whether required evidence remains retrievable as new episodes arrive. Storing earlier evidence is insufficient if later retrieval cannot select it when needed. With `gpt-4.1-nano`, the four baselines obtain State F1 scores of 23.39%–23.83%, and only 43.75%–48.85% of transitions starting from the highest mapped judge grade retain that grade (Tables 3 and 4). These failures motivate examining evidence access during the process. MemStream pairs reference answers with supporting evidence at successive information states. Comparing retrieval across these states can show when previously retrieved, still-valid evidence is lost or newly required evidence is missed, providing a diagnostic dimension absent from final-state answer scores.

4.2 Construction and Evaluation Protocol

Figure 3 summarizes the construction and evaluation protocol. We audit 3,947 source question–answer (QA) pairs from Mem-Gallery and H2HMem (Bei et al., 2026; Zhu et al., 2026), identify 779 evolution-related questions, and annotate 1,603 information states with reference answers and supporting evidence. Each question receives one of seven categories of information evolution. We retain the original episode and timestamp order.

For question q_p, let p index its question trajectory and L_p be its number of checkpoints. At checkpoint ℓ, $n_{p,\ell}$ episodes have arrived, $s_{p,\ell}$ is the information state, $a_{p,\ell}$ its reference answer, and $\mathcal{E}_{p,\ell}$ its supporting episode set. The trajectory is

$$\mathcal{T}_p = \left(q_p, \left(n_{p,\ell}, s_{p,\ell}, a_{p,\ell}, \mathcal{E}_{p,\ell}\right)_{\ell=1}^{L_p}\right). \tag{5}$$

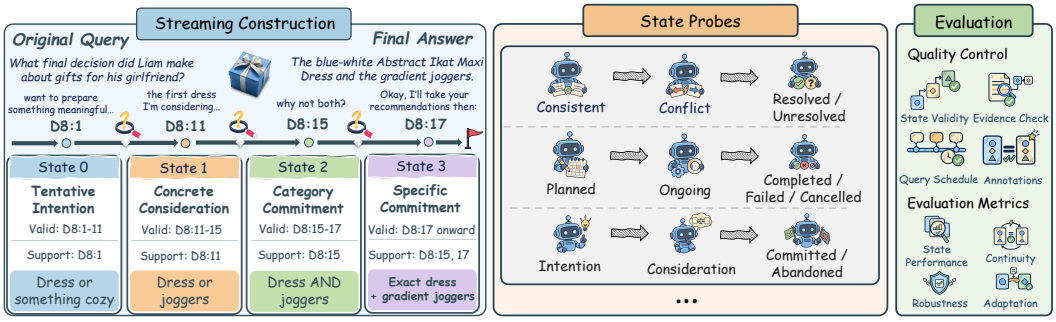

Figure 3: **Overview of MemStream.** Each question is evaluated at successive conversation prefixes with state-specific reference answers and supporting evidence. New evidence denotes support newly required relative to the preceding state. Process metrics assess answer correctness, continuity, and access to still-valid evidence.

The system can access only $\mathcal{D}_{1:n_{p,\ell}}$, the available prefix defined in Section 3.1; future episodes are hidden. We delay each checkpoint probe within the interval in which its reference answer remains valid, allowing unrelated episodes to arrive after its supporting evidence. Validation checks that evidence belongs to the visible prefix, supports the state-specific answer, and yields the original source answer at the final state. The annotations require no AMBER graph or association labels.

4.3 Process-Level Metrics

To examine evidence access, we compare the retrieved episode set $\widehat{\mathcal{E}}_{p,\ell}$ with $\mathcal{E}_{p,\ell}$. Support R@10 measures current evidence coverage; Retained R@10 measures coverage of earlier evidence that remains necessary. Recall of newly required evidence measures adaptation to the current state's evidence requirements.

State F1 averages token-level F1 between generated and reference answers over all checkpoints; State LLM-J averages their reference-guided judge scores. Let $J_{p,\ell} \in \{0, 0.5, 1\}$ be the mapped judge score at checkpoint ℓ for question p. Its highest grade, $J_{p,\ell} = 1$, corresponds to a raw score of 0.75 or 1. Let $b_{p,\ell} = \mathbf{1}[J_{p,\ell} = 1]$ indicate a top-grade answer. Each answer is judged against its current reference, allowing valid answers to change across states.

For P evaluated questions, the adjacent transitions form $\mathcal{A} = \{(p, \ell) : 1 \le p \le P,\ 1 \le \ell < L_p\}$. *Strict Continuity* (SC) measures the proportion with both answers receiving the top grade. *Correct Retention* (CR) conditions on the earlier answer receiving this grade:

$$\mathrm{SC} = \frac{\sum_{(p,\ell)\in\mathcal{A}} b_{p,\ell} b_{p,\ell+1}}{|\mathcal{A}|}, \qquad \mathrm{CR} = \frac{\sum_{(p,\ell)\in\mathcal{A}} b_{p,\ell} b_{p,\ell+1}}{\sum_{(p,\ell)\in\mathcal{A}} b_{p,\ell}}. \tag{6}$$

Soft Continuity averages $\min(J_{p,\ell}, J_{p,\ell+1})$ over $\mathcal{A}$, retaining partial credit. *Degradation Rate* is $1 - \mathrm{CR}$, the probability of receiving a lower grade at the next checkpoint given a top-grade answer.

Paired evidence and answer records allow checking whether a decline in answer quality coincides with missing support. This evaluates the combined effect of construction and retrieval on evidence access. Definitions of the trajectory-level metrics, evidence measures, and annotation categories appear in Appendix F.

5 Experiments

5.1 Experimental Settings

Datasets. Mem-Gallery (Bei et al., 2026) contains 1,711 questions across 20 scenarios and 240 sessions of user-assistant conversations. H2HMem (Zhu et al., 2026) contains 2,236 questions from dyadic and multi-party conversations with asynchronous multimodal evidence. MemStream (Section 4) provides 1,603 annotated information states for 779 evolution-related questions drawn from these two benchmarks.

Table 1: Main results on Mem-Gallery, H2HMem, and MemStream. Scores are percentages (%) on a 0–100 scale and macro-averaged across scenarios with equal weights. LLM-J uses `gpt-4o-mini`. Best results are bold and runners-up are underlined. Each Δ row reports the signed percentage-point difference between AMBER and the strongest reported baseline: $\Delta = s_{\mathrm{AMBER}} - s_{\mathrm{best}}$, where s is the displayed percentage score.

Method	Mem-Gallery				H2HMem				MemStream			
	F1	BLEU-1	EM	LLM-J	F1	BLEU-1	EM	LLM-J	F1	BLEU-1	EM	LLM-J
Answer backbone: `gpt-4.1-nano`												
MuRAG	59.06	53.13	35.89	75.60	29.06	23.43	12.88	53.67	23.83	16.56	1.75	55.02
NGM	58.34	52.43	37.17	74.02	27.89	22.75	12.97	49.37	23.77	16.67	2.68	51.59
AUGUSTUS	53.60	47.22	31.33	70.66	28.13	22.93	13.06	50.58	23.81	16.71	2.43	51.03
UniversalRAG	43.76	38.77	25.25	54.21	25.05	20.51	11.67	45.04	23.39	16.39	1.56	51.87
AMBER (ours)	**62.60**	**57.63**	**41.15**	**75.98**	**30.59**	**24.74**	**13.51**	**55.43**	**29.05**	**21.35**	**3.74**	**57.02**
Δ	**+3.54**	**+4.50**	**+3.98**	**+0.38**	**+1.53**	**+1.31**	**+0.45**	**+1.76**	**+5.22**	**+4.64**	**+1.06**	**+2.00**
Answer backbone: `gpt-5.4-mini`												
MuRAG	72.61	68.23	54.70	84.13	34.74	29.04	17.22	63.19	29.87	22.99	8.23	64.22
NGM	69.73	65.25	52.48	81.09	33.62	28.31	17.44	59.12	28.80	22.04	7.55	60.42
AUGUSTUS	67.49	62.92	49.50	78.93	32.79	27.43	16.14	58.54	27.62	20.44	5.86	60.64
UniversalRAG	53.97	49.42	38.57	63.30	28.73	24.11	14.13	51.79	29.44	22.39	7.74	62.13
AMBER (ours)	**74.18**	**69.94**	**56.75**	**85.30**	**37.70**	**31.73**	**18.56**	**66.23**	**31.11**	**24.30**	**8.61**	**64.66**
Δ	**+1.57**	**+1.71**	**+2.05**	**+1.17**	**+2.96**	**+2.69**	**+1.12**	**+3.04**	**+1.24**	**+1.31**	**+0.38**	**+0.44**

Baselines. We use the Mem-Gallery implementations of four baselines. **MuRAG** (Chen et al., 2022) retrieves independent multimodal records by semantic similarity. **Neural Graph Memory (NGM)** (Fisher, 2025) uses a graph for similarity-guided retrieval. **AUGUSTUS** (Jain et al., 2025) connects multimodal records through semantic concepts. **UniversalRAG** (Yeo et al., 2025) selects between document retrieval, image retrieval, and answering without retrieval.

Implementation. We use `gpt-4.1-nano` and `gpt-5.4-mini` as answer backbones. All methods receive the same histories and queries with a retrieval budget of $K = 10$ episodes. Semantic retrieval uses the frozen `gme-Qwen2-VL-2B-Instruct` encoder, 1,536-dimensional embeddings, and cosine similarity. All LLM-assisted components use `gpt-4.1-nano` with temperature 0. AMBER uses $\lambda = 0.5$ for chain scoring. The evidence context is limited to 4,096 tokens and 20 images by the generation pipeline. Memory construction and retrieval have no access to benchmark annotations or future episodes.

Metrics. Following the Mem-Gallery implementation, we report normalized token F1 with Porter stemming, smoothed BLEU-1, exact match (EM), and LLM-as-a-Judge (LLM-J). LLM-J uses `gpt-4o-mini` to compare predictions with reference answers, mapping raw grades $(0, 0.25, 0.5, 0.75, 1)$ to $(0, 0.5, 0.5, 1, 1)$ before aggregation and process evaluation. Main answer scores are averaged within each scenario and then equally across scenarios. Reported score changes are absolute percentage-point differences. Under the ten-episode retrieval budget, Support R@10 and Retained R@10 measure recall of current supporting evidence and still-required earlier evidence, respectively. Hit@10 records whether any current support is retrieved. Normalized discounted cumulative gain (nDCG@10) (Järvelin & Kekäläinen, 2002) measures ranking quality. P–R H is the harmonic mean of retained recall and recall of newly required evidence. Evidence and process metrics follow Section 4.3 and Appendix F.

5.2 Main Results

Table 1 shows that AMBER ranks first in all 24 combinations of dataset, answer backbone, and metric. Averaged equally across the six dataset–backbone settings, its gains over the strongest baseline per metric are 2.68% in F1, 2.69% in BLEU-1, 1.51% in EM, and 1.47% in LLM-J. Mean F1 gains across the two backbones are 2.56%, 2.25%, and 3.23% on Mem-Gallery, H2HMem, and MemStream, respectively. Within each setting, the answer backbone and generation procedure are shared, so these

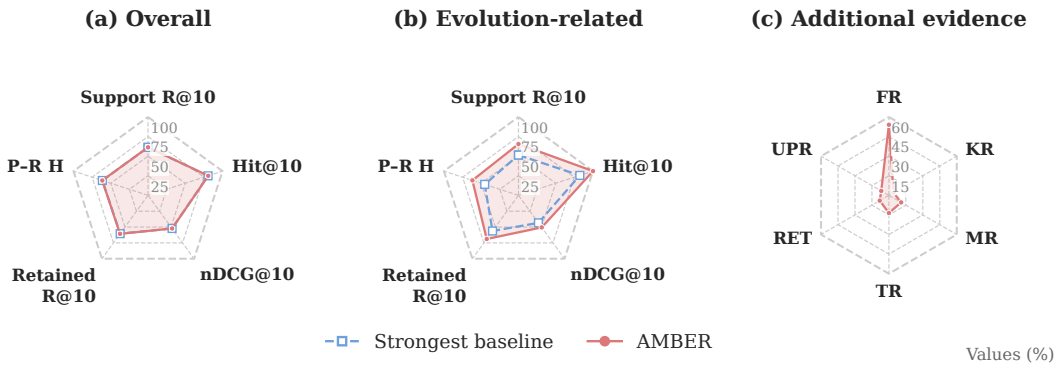

Figure 4: **Evidence retrieval across evaluation sets and query types.** (a,b) Overall and evolution-related retrieval use a shared 0–100% scale and the strongest baseline per metric; the subset follows the MemStream screening criteria. (c) Shares of the 74 additionally retrieved pieces of evidence (0–60% scale), showing query types with at least two items. Full results and query-type definitions appear in Tables 5–7.

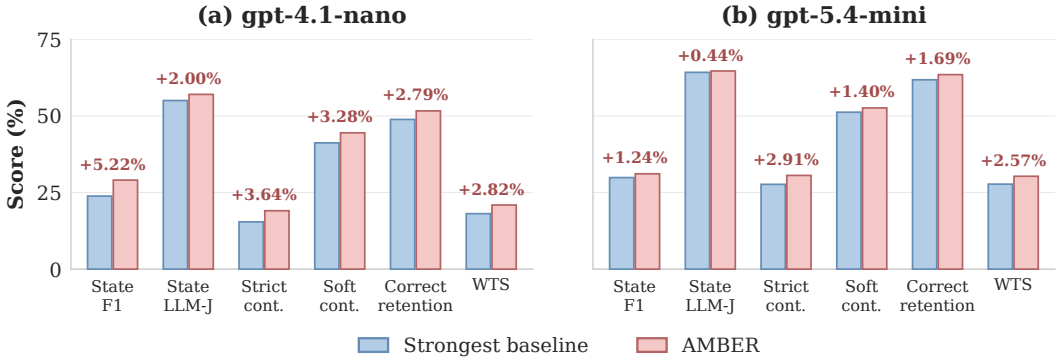

Figure 5: **MemStream process performance under two answer backbones.** Bar pairs compare AMBER with the strongest baseline for each of the six metrics; annotations report score differences (%). Both panels use a shared 0–75% scale, with higher values better for all displayed metrics. WTS denotes Whole-Trajectory Success. Complete results, including Hidden Earlier Failure (HEF) and Degradation Rate, appear in Tables 3 and 4.

comparisons assess the evidence supplied by each memory system. The following analyses examine whether improved answer quality is accompanied by better access to required evidence and greater continuity across information states.

5.3 Performance Analysis

Retrieval Performance. Figure 4 shows improvements across all five retrieval metrics. Gains over the strongest baseline range from 0.02% to 0.33% overall and increase to 7.17%–17.65% on the evolution-related subset. Support R@10 improves by 14.05% and Retained R@10 by 12.57% on this subset, indicating better access to both current support and still-required earlier evidence. On evolution-related queries with association activation, AMBER additionally retrieves 74 of the 388 pieces of supporting evidence missed by similarity retrieval (19.07%). Factual Retrieval (FR) accounts for 54.05% of the additionally retrieved evidence. This breakdown groups evidence by benchmark query type and does not measure the contribution of an individual association type.

Information Evolution Performance. AMBER leads all nine process metrics under both answer backbones. Figure 5 shows its advantage across state-level answer quality, adjacent-state continuity, and whole-trajectory success. Compared with the strongest baseline, Strict Continuity improves by 3.64% and 2.91%, and Whole-Trajectory Success by 2.82% and 2.57%, for `gpt-4.1-nano`

and `gpt-5.4-mini`, respectively. HEF decreases by 8.20% and 4.49%, while Degradation Rate decreases by 2.79% and 1.69%. These improvements show more consistent answers as the available information changes, alongside the retrieval gains above.

5.4 Ablation and Hyperparameter Analysis.

Table 2(a) reports ablations from an independent H2HMem rerun, including a fresh run of the full configuration. We test the two retrieval mechanisms in Section 3.4. Without *matching through a specific association*, seed episodes can expand along paths of all association types. Without *searching and ranking with a generated anchor*, retrieval omits anchor-based matching and ranks chains by their strongest seed similarity alone. These variants reduce answer F1 by 0.47% and 0.27%, respectively, relative to Full in the same rerun. Full also outperforms Similarity-only, which selects independent episodes by semantic similarity, by 0.76%, 0.90%, and 0.90% in answer F1, BLEU-1, and EM, respectively.

Panels (b) and (c) vary the seed and final episode budgets. The seed budget K_s controls entry points before path expansion; varying it from 3 to 20 gives score ranges of 0.30%, 0.29%, and 0.24% in F1, BLEU-1, and EM on Mem-Gallery. The final budget K caps the episodes selected after chain ranking. Increasing K from 3 to 10 improves these answer metrics by 5.81%, 6.28%, and 5.97%, respectively. Increasing it from 10 to 20 changes them by -0.01%, $+0.03\%$, and -0.88%, while doubling the permitted episode budget. We use $K_s = 10$ and $K = 10$ as defaults.

Table 2: Retrieval ablation and hyperparameter analysis. All scores are percentages (%). Panel (a) reports independently rerun H2HMem configurations.

Configuration	F1	BLEU-1	EM
(a) Ablation on H2HMem			
Similarity-only	29.20	23.61	13.11
w/o specific association	29.49	23.96	13.44
w/o anchor generation	29.69	24.14	13.52
AMBER (Full)	**29.96**	**24.51**	**14.01**
(b) Seed Top-K_s on Mem-Gallery			
$K_s = 3$	71.50	67.07	54.06
$K_s = 5$	71.57	67.14	54.24
$K_s = 10$ (default)	71.41	66.90	**54.30**
$K_s = 20$	**71.71**	**67.19**	**54.30**
(c) Final Budget on Mem-Gallery			
$K = 3$	65.60	60.62	48.33
$K = 5$	68.80	64.44	52.95
$K = 10$ (default)	**71.41**	66.90	**54.30**
$K = 20$	71.40	**66.93**	53.42

5.5 Qualitative Analysis

In the examined cases, MuRAG retrieves relevant episodes but misses required evidence. AMBER generates anchors about *human oversight*, *making bread in the baking course*, *applications of AI in cybersecurity*, and *Leo's chosen natural law* for the temporal, procedural, complementary, and conflicting cases, respectively. Each query anchor, together with the selected association type, determines which stored paths are expanded from similarity seeds. Seeds on the same path form one candidate evidence chain, ranked by its strongest seed similarity and agreement with the query anchor. Within the same Top-10 episode budget, these chains include the earlier discussion, subsequent baking steps, additional applications, and final-choice evidence missed by MuRAG. Evidence assembly retains chain order and includes the association type and anchor to contextualize the selected episodes for answer generation. These intermediate steps improve evidence coverage in the examined cases, although full coverage does not ensure that every supported detail appears in the answer. Appendix G presents the corresponding questions, anchors, retrieval decisions, evidence chains, and generated answers.

6 Conclusion

We presented AMBER, a multimodal memory framework that links episodes through explicit associations and retrieves query-relevant evidence chains. We also introduced MemStream to evaluate evidence access and answer continuity as information evolves. Across three benchmarks and two backbones, AMBER improves average F1 by 2.68% over the strongest baselines. On cross-episode queries, Support R@10 and Retained R@10 rise by 14.05% and 12.57%. These results highlight the value of our rich association in bridging episodes for multimodal memory.

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.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16470–16489. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-Gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 40750–40784. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330. Springer, 2001.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026. doi: 10.48550/arXiv.2605.06527. URL https://arxiv.org/abs/2605.06527.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.375. URL https://aclanthology.org/2022.emnlp-main.375/.

Yi Chen, Wei Wang, Ziyang Liu, and Xuemin Lin. Keyword search on structured and semi-structured data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 1005–1010, 2009.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. doi: 10.48550/arXiv.2504.19413. URL https://arxiv.org/abs/2504.19413.

Curtis Dyreson, Fabio Grandi, Wolfgang Käfer, Nick Kline, Nikos Lorentzos, Yannis Mitsopoulos, Angelo Montanari, Daniel Nonen, Elisa Peressi, Barbara Pernici, et al. A consensus glossary of temporal database concepts. *ACM SIGMOD Record*, 23(1):52–64, 1994.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130. URL https://arxiv.org/abs/2404.16130.

Matthew Fisher. Neural Graph Memory: A structured approach to long-term memory in multimodal agents, 2025. URL https://www.researchgate.net/publication/394440420_Neural_Graph_Memory_A_Structured_Approach_to_Long-Term_Memory_in_Multimodal_Agents. Preprint.

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*, 2024. URL https://arxiv.org/abs/2405.14831.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025.

Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. BLINKS: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, pp. 305–316, 2007. doi: 10.1145/1247480.1247516.

Aidan J. Horner, James A. Bisby, Daniel Bush, Wen-Jing Lin, and Neil Burgess. Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, 6:7462, 2015. doi: 10.1038/ncomms8462.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. AUGUSTUS: An LLM-driven multimodal agent system with contextualized user memory. *arXiv preprint arXiv:2510.15261*, 2025. URL https://arxiv.org/abs/2510.15261.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.06036. Also available as arXiv:2606.06036.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. MAGMA: A multi-graph based agentic memory architecture for AI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36848–36865. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1709.

Dingyi Kang, Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. V-Mem: Modality-routed retrieval for long-term multimodal agentic memory. *arXiv preprint arXiv:2608.01543*, 2026. URL https://arxiv.org/abs/2608.01543.

LangChain AI. LangMem: Long-term memory for LLM agents. https://github.com/langchain-ai/langmem, 2026. Accessed: 2026-09-17.

Shiyang Li, Yifan He, Haonan Guo, Xingwei Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xiaoye Qu, Yang Li, Wanli Ouyang, et al. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12758–12786, 2024.

Leonid Libkin, Wim Martens, and Domagoj Vrgoč. Querying graphs with data. *Journal of the ACM*, 63(2):1–53, 2016. doi: 10.1145/2850413.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Jia Nan, Weizhi Ma, Wenhui Wu, and Yanyan Chen. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341*, 2025. URL https://arxiv.org/abs/2508.03341.

Thomas Neumann and Gerhard Weikum. RDF-3X: A RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.

Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. PersonaVLM: Long-term personalized multimodal LLMs. *arXiv preprint arXiv:2604.13074*, 2026. URL https://arxiv.org/abs/2604.13074. Accepted by CVPR 2026.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8631–8661. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.435. URL https://aclanthology.org/2025.naacl-long.435/.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL https://aclanthology.org/L08-1093/.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, December 2014. doi: 10.1162/COLI_a_00204. URL https://aclanthology.org/J14-4007/.

Peter Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

Richard T. Snodgrass. *The TSQL2 Temporal Query Language*. Springer Science & Business Media, 2012.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29935–29951. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1496. URL https://aclanthology.org/2026.findings-acl.1496/.

Jingyu Sun, Yuyang Xue, Mingyang Li, Zhengtao Yao, Jiachen Li, Yang Cui, Wenhao Cai, Haozhe Liu, Fangying Wang, Magdalene Katharina Montgomery, Syed Murtuza Baker, and Hongpeng Zhou. TrajWiki: Source-grounded memory trajectories for long-horizon dialogue agents. *arXiv preprint arXiv:2608.00967*, 2026. URL https://arxiv.org/abs/2608.00967.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. MemConflict: Evaluating long-term memory systems under memory conflicts. *arXiv preprint arXiv:2605.20926*, 2026. doi: 10.48550/arXiv.2605.20926. URL https://arxiv.org/abs/2605.20926.

Thanh Tran, Haofen Wang, Sebastian Rudolph, and Philipp Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped RDF data. In *Proceedings of the 25th IEEE International Conference on Data Engineering*, pp. 405–416, 2009.

Maolin Wang, Yu Wang, Zichun Liu, Baiyuan Qiu, Chenbin Zhang, Jiguang Shen, Haoran Yang, and Hao Miao. TRACE: State-aware query processing over temporal evidence graphs for conversational data. *arXiv preprint arXiv:2607.00339*, 2026a. URL https://arxiv.org/abs/2607.00339.

Wenxuan Wang, Haoyu Sun, Fukuan Hou, Mingyang Song, Weinan Zhang, Yu Cheng, and Yang Yang. SubtleMemory: A benchmark for fine-grained relational memory discrimination in long-horizon AI agents. *arXiv preprint arXiv:2606.05761*, 2026b. URL https://arxiv.org/abs/2606.05761.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-Memory: Benchmarking LLM agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2026. URL https://arxiv.org/abs/2511.20857.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024. URL https://arxiv.org/abs/2410.10813.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5180–5197, 2022. doi: 10.18653/v1/2022.acl-long.356. URL https://aclanthology.org/2022.acl-long.356/.

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*, volume 38, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. UniversalRAG: Retrieval-augmented generation over multiple corpora with diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*, 2025. URL https://arxiv.org/abs/2504.20734.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Shiping Zhu, Yibo Yang, Zhengyang Wang, Tiancheng Shen, Dandan Guo, and Ming-Hsuan Yang. H2HMem: A multimodal memory benchmark for agents in human-human interactions. *arXiv preprint arXiv:2606.09461*, 2026. doi: 10.48550/arXiv.2606.09461. URL https://arxiv.org/abs/2606.09461.

Lei Zou, Jinghui Mo, Lei Chen, M. Tamer Özsu, and Dongyan Zhao. gStore: Answering SPARQL queries via subgraph matching. *Proceedings of the VLDB Endowment*, 4(8):482–493, 2011.

Appendix Contents

A Association Taxonomy **15**

A.1 Definition of Association Types 15

A.2 Multiple Association Types and Ambiguous Cases 15

B Detailed Results **15**

C Prompts **15**

C.1 Association Construction 17

C.2 Query-Conditioned Activation 18

C.3 Association-Aware Evidence Assembly 18

C.4 Answer Generation 19

C.5 LLM-as-a-Judge 20

D Implementation Details **20**

D.1 Online Association Construction 20

D.2 Query-Conditioned Activation 21

D.3 Evidence Chain Ranking 22

D.4 Association-Aware Evidence Assembly 22

E Generalization and Robustness **22**

E.1 Task Generalization across Sources and Backbones 22

E.2 Robustness to Memory Perturbations 23

E.3 Label-Preserving and Counterfactual Perturbations 23

F MemStream Annotation and Metrics **24**

F.1 Construction Pipeline 24

F.2 Overall Statistics 24

F.3 Answer Metric Definitions 24

F.4 Evidence Access and Retention 25

G Case Studies **25**

G.1 Temporal Association 26

G.2 Procedural Association 26

G.3 Complementary Association 27

G.4 Conflicting Association Mode 28

H Limitations **29**

I Reproducibility and AI Use Statement **29**

A Association Taxonomy

A.1 Definition of Association Types

The four association types distinguish temporal order, procedural dependency, information supplementation, and incompatible assertions for evidence retrieval. The Penn Discourse TreeBank (PDTB) distinguishes Temporal, Contingency, Comparison, and Expansion relations (Prasad et al., 2008). This provides a motivation for distinguishing semantic functions, not a proof that AMBER's four types exhaust the relations between memory episodes. The units and operational definitions differ from discourse annotation, so the correspondence is functional rather than a conversion between label sets. In particular, PDTB's Comparison includes contrasts between statements that can both be true, whereas AMBER's conflicting association requires a contradiction. Later analysis of PDTB discusses both concurrent relations and gaps in its sense inventory (Prasad et al., 2014).

The temporal type represents explicit ordering or time references between episodes. A timestamp on an episode alone does not make every association temporal. The present definition uses *precedes*; it does not claim to implement the full range of temporal relations in discourse annotation. The procedural type represents a step that enables another step within the same procedure. It has a functional connection to contingency-style dependencies, but does not include every causal, conditional, or pragmatic relation in PDTB. Mere chronological succession is insufficient to establish a procedural dependency. Here, *procedural* describes a relation between episode contents, not procedural memory as a psychological memory system. The complementary type represents an episode that extends or fills in information about an anchor without contradicting the information it supplements. It is related to the elaborative function of discourse expansion, but does not equate all expansion relations with new evidence. SubtleMemory also studies complementary memories, alongside nuanced and contradictory relations (Wang et al., 2026b); these categories motivate relational discrimination without establishing the coverage of AMBER's inventory. The conflicting type requires incompatible assertions about the same anchor; different descriptions or values can be compatible when they apply to different attributes or contexts. A later assertion supersedes an earlier one when it resolves the conflict by replacing that assertion. Temporal order alone does not establish either contradiction or supersession.

A.2 Multiple Association Types and Ambiguous Cases

For episodes e_i and e_j, let $\mathcal{R}_{ij} \subseteq \mathcal{R}$ denote the set of supported association types. This set may contain more than one type, or be empty. For example, a pair of steps may have both an explicit temporal order and an enabling dependency. Such overlap is compatible with the independent construction decisions for each association type described in the main text and with observations of concurrent discourse relations (Prasad et al., 2014).

Multiple supported relations are distinct from uncertainty about which relation holds: shared vocabulary alone does not establish a dependency, and a difference in wording alone does not establish a contradiction. At retrieval time, AMBER still selects exactly one mode according to the query. This is a retrieval policy, not an assumption that stored association types are mutually exclusive. Questions requiring joint traversal across several types remain a limitation of this policy.

B Detailed Results

Tables 3 and 4 report state-level and process-level answer results. All entries are percentages; larger values are better except for HEF and Degradation Rate. State averages include all evaluated checkpoints. Conditional metrics use the denominators specified in Appendix F.

C Prompts

AMBER uses LLM prompts for association construction and query-conditioned activation, then assembles the selected evidence for answer generation. The templates below are condensed renderings of the implemented prompts and message layout, with field names normalized to the manuscript's

Table 3: State and question-trajectory answer quality on MemStream (%). Query-Macro averages questions equally; Whole-Trajectory Success (WTS) requires the highest mapped judge grade ($J = 1$) at every state. Hidden Earlier Failure (HEF) is restricted to multi-state trajectories with $J = 1$ at the final state.

Method	State F1	State LLM-J	Query-Macro LLM-J	WTS	HEF↓
Answer backbone: `gpt-4.1-nano`					
MuRAG	23.83	55.02	55.05	18.10	59.12
NGM	23.77	51.59	51.76	16.17	58.22
AUGUSTUS	23.81	51.03	51.05	16.05	59.24
UniversalRAG	23.39	51.87	51.19	14.89	57.33
AMBER	**29.05**	**57.02**	**56.88**	**20.92**	**49.13**
Answer backbone: `gpt-5.4-mini`					
MuRAG	29.87	64.22	63.76	27.73	53.33
NGM	28.80	60.42	59.58	24.52	52.05
AUGUSTUS	27.62	60.64	59.87	20.80	60.55
UniversalRAG	29.44	62.13	61.22	26.06	48.93
AMBER	**31.11**	**64.66**	**64.26**	**30.30**	**44.44**

Table 4: Answer continuity across adjacent MemStream checkpoints (%). Correct Retention conditions on the highest mapped grade ($J = 1$) at the earlier state; Degradation Rate is its complement.

Method	Strict Continuity	Soft Continuity	Correct Retention	Degradation Rate↓
Answer backbone: `gpt-4.1-nano`				
MuRAG	15.41	41.20	48.85	51.15
NGM	13.59	37.86	43.75	56.25
AUGUSTUS	12.99	37.44	45.53	54.47
UniversalRAG	13.83	39.38	44.88	55.12
AMBER	**19.05**	**44.48**	**51.64**	**48.36**
Answer backbone: `gpt-5.4-mini`				
MuRAG	27.67	51.21	61.79	38.21
NGM	24.03	48.54	58.41	41.59
AUGUSTUS	21.00	47.39	51.95	48.05
UniversalRAG	27.06	50.24	60.76	39.24
AMBER	**30.58**	**52.61**	**63.48**	**36.52**

Table 5: Overall evidence retrieval (%). R@10 denotes recall of current support (Support) or still-required earlier evidence (Retained) under a ten-episode budget. Hit@10 indicates whether any current support is retrieved; nDCG denotes normalized discounted cumulative gain. P–R H is the harmonic mean of retained recall and recall of newly required evidence.

Method	Support R@10	Hit@10	nDCG@10	Retained R@10	P–R H
MuRAG	61.23	80.54	52.24	60.44	61.30
NGM	47.97	67.56	37.20	46.19	48.07
AUGUSTUS	49.90	74.61	43.39	49.91	49.90
UniversalRAG	56.72	76.23	48.34	57.20	56.67
AMBER	**61.30**	**80.79**	**52.26**	**60.77**	**61.35**

Table 6: Evidence retrieval on the evolution-related subset (%). Metric abbreviations follow Table 5. The strongest baseline is selected separately for each metric.

Method	Support R@10	Hit@10	nDCG@10	Retained R@10	P–R H
Strongest baseline	51.41	82.35	43.39	56.28	45.28
AMBER	**65.46**	**100.00**	**50.56**	**68.85**	**61.80**

Table 7: Additional evidence retrieval by query type across 346 evolution-related information states with association activation. MG denotes Mem-Gallery, H2H denotes H2HMem, and Both pools identically named categories from the two datasets. Missed counts pieces of supporting evidence absent from similarity retrieval; additionally retrieved counts those subsequently retrieved by AMBER. Share is the additionally retrieved count divided by 74; additional retrieval rate is the additionally retrieved count divided by the missed count. A dash denotes an undefined rate when no evidence was missed.

Type	Meaning	Source	States	Missed	Additionally retrieved	Share (%)	Additional retrieval rate (%)
FR	Factual Retrieval	MG	63	76	40	54.05	52.63
KR	Knowledge Resolution	Both	36	47	3	4.05	6.38
MR	Multi-entity Reasoning	MG	10	40	8	10.81	20.00
TR	Temporal Reasoning	Both	109	107	10	13.51	9.35
MCR	Multimodal Causal Inference	H2H	13	13	1	1.35	7.69
RET	Reference & Evolution Tracking	H2H	54	46	6	8.11	13.04
CRR	Cross-modal Related Retrieval	H2H	10	9	1	1.35	11.11
UPR	Unimodal Precise Recall	H2H	46	50	5	6.76	10.00
Overall	–	MemStream	346	388	74	100.00	19.07

terminology and each anchor shown as one textual description. Angle-bracketed fields denote variable inputs. Answer generation and judging follow the Mem-Gallery implementation (Bei et al., 2026).

C.1 Association Construction

The controller is called independently for each association type. It receives the incoming episode's text and timestamp, textual cue scores and matched patterns for all four types, and candidate stored-path summaries. Each summary contains an identifier, association type, anchor, episode count, and ordered timestamps. These summaries supply the existing contexts to which the new episode may belong; they do not contain the earlier episodes' full text or images.

Association construction: condensed prompt

Instruction

Decide whether to extend an existing path, initialize a path, or make no update for the target association only. `EXTEND` requires exactly one advertised candidate path identifier and a null new anchor. `INITIALIZE` requires a nonempty new anchor and a null path identifier. `NONE` requires both fields to be null. For temporal associations, use valid timestamps to verify chronology. Return only one JSON object with the specified output fields.

Input fields

Target association: `<temporal / procedural / complementary / conflicting>`
Incoming episode: `<text, timestamp>`
Textual signals: `<type, score, matched patterns>` for each type
Candidate paths: `<path ID, type, anchor, episode count, ordered timestamps>`

Output fields

`action`: `EXTEND`, `INITIALIZE`, or `NONE`; `path_id`: integer or null; `anchor`: description or null; `confidence`: number in $[0, 1]$; `reason`: short rationale.

Validation enforces the action-specific fields and confidence threshold. An invalid or low-confidence decision leaves that association type unchanged; the episode is still stored. Decisions for the other types remain independent.

C.2 Query-Conditioned Activation

The activator receives the question text and the four textual signal summaries. It selects one retrieval mode and generates an anchor for an association mode. It does not receive retrieved evidence or generate an answer. Query images remain available to semantic retrieval and the answer model.

Query-conditioned activation: condensed prompt

Instruction

Select exactly one retrieval mode: temporal, procedural, complementary, conflicting, `SIMILARITY_ONLY`, or `NO_RETRIEVAL`. Use `SIMILARITY_ONLY` for direct semantic or visual memory retrieval. Use `NO_RETRIEVAL` only when the current query input is explicitly sufficient, and set `current_input_sufficient` to true. Do not answer the question. Return one JSON object.

Input fields

Question: `<question text>`
Textual signals: `<type, score, matched patterns>` for each type

Output fields

`mode`: one allowed mode; `anchor`: description for an association mode, otherwise null; `confidence`: number in $[0, 1]$; `reason`: short rationale; `current_input_sufficient`: boolean.

For the baking question, an illustrative activation is procedural with anchor *making bread in the baking course*. This anchor describes the process being sought; it does not supply the missing steps. Invalid or low-confidence activation falls back to semantic retrieval alone.

C.3 Association-Aware Evidence Assembly

Assembly converts the selected chains $\mathbb{C}_t^*$ into the multimodal context $\mathcal{X}_t$ consumed by the answer model. Its purpose is to carry the associations used in retrieval into the evidence presentation. The association type states how the episodes are connected, the anchor states what they concern, and the path position preserves their order. These descriptors accompany the original evidence rather than replacing it with a generated summary. Assembly is deterministic and requires no additional LLM call.

From selected chains to evidence blocks. The organizer visits chains in their selected ranking order and episodes in each chain's stored order. For a selected chain containing multiple episodes, each episode receives a descriptor with its association type, source path identifier, anchor, and position in that path. Positions refer to the original stored path even when the budget selects only a contiguous window. Thus, a cropped chain is not renumbered as though it were the complete path. Repeated episodes are emitted only at their first occurrence, and singletons are supplied without a multi-episode association descriptor.

Each evidence block retains the episode text, available captions, timestamp, and source identifier. Its original image, when available, follows as a multimodal content block. The text includes an image reference rather than encoded image bytes. Consequently, the answer model receives both the association context and the visual evidence associated with that episode. The organizer does not add unobserved steps, resolve contradictions, or perform another retrieval pass.

Association-aware evidence context: normalized message layout

System message

`<shared Mem-Gallery answer instructions>`

User message: retrieved evidence

The retrieved memory contents are as follows:

```
[AMBER association=<type> path=<ID>
       anchor=<description> position=<index>/<path length>]
text:  <original episode text>
image:  <image reference, if present>
```

```
caption:  <available caption, if present>
timestamp:  <episode timestamp, if present>
source ID: <episode source identifier, if present>
```

Original image content is inserted here as a separate multimodal block, if available. Repeat the evidence block in selected chain order, retaining episode order within each chain.

User message: question

`<shared question instructions>`
The current question is as follows:
`<question text> <task-specific format constraint>`
The question image follows, if present.

Baking example. For the introductory question about experiences and steps in a baking course, the anchor *making bread in the baking course* connects the selected classroom episodes to the concrete preparation and baking steps. An illustrative evidence context follows the order below. The descriptions are shortened for presentation; the executed context retains the selected episodes' original text and available images.

Illustrative procedural evidence chain

Association: procedural **Anchor:** making bread in the baking course

`D2:1--D2:3`: enter the classroom; learn about tools and ingredients
`D2:4`: weigh the flour
`D2:5`: add ingredients, stir, and knead the dough
`D2:6`: ferment the dough
`D2:7`: shape the fermented dough
`D2:8`: bake the bread until golden brown
`D2:9`: take out the finished bread

This presentation keeps the specific actions connected to the course that the question names. A complementary chain similarly keeps supplements under a common anchor, while temporal and conflicting chains retain the ordering and source contents needed to interpret earlier and later assertions. An association label describes the retrieved connection; the answer must still be grounded in the episode contents.

C.4 Answer Generation

The shared Mem-Gallery system prompt requests concise but complete answers grounded in the multimodal conversation and instructs the model to rely on the most recent information when an event is updated. It describes the benchmark's question types and prohibits unsupported answers. AMBER inserts the assembled context into this shared protocol; it does not add a separate association-specific reasoning instruction.

Shared question prompt: condensed rendering

Answer the question about the conversation between `<speaker A>` and `<speaker B>` concisely using the memory content. Provide only the answer, without an introductory phrase such as "answer:". For a date or time question, provide a specific date or time whenever possible. Keep the answer concise but complete.

The current question is as follows:
`<question> <format constraint>`

The format constraint depends on the source task. Conflict detection requests "Yes." or "No."; visual-centric search requests image identifiers in ascending order, separated by commas; answer refusal requests "Not mentioned." when the requested information is absent. These constraints govern output format, while the evidence context determines what information is available to answer the question.

C.5 LLM-AS-A-JUDGE

The inherited judge prompt receives the question, reference answer, and model answer. It requests a JSON object containing `score` and `reasoning`. Its five raw grades are 0 for an incorrect or missing answer, 0.25 for a tangential answer missing the core information, 0.5 for a partially correct answer, 0.75 for an accurate answer with minor imperfections, and 1 for a precise and complete answer without hallucinations. For yes/no questions, the rubric explicitly checks polarity.

Judge input and output

Question: `<question>`
Ground Truth: `<reference answer>`
Assistant Answer: `<generated answer>`
Output: `{"score": <0, 0.25, 0.5, 0.75, or 1>, "reasoning": "<short explanation>"}`

The evaluation implementation maps these raw grades to 0, 0.5, 0.5, 1, and 1, respectively. Reported LLM-J and judge-based MemStream process metrics use the mapped scores. At a MemStream checkpoint, generation uses only the available conversation prefix, and judging uses that checkpoint's reference answer. State-specific reference answers and annotated supporting evidence are evaluation targets, not inputs to retrieval or answer generation.

D IMPLEMENTATION DETAILS

D.1 ONLINE ASSOCIATION CONSTRUCTION

Candidate paths and pattern signals. Before storing e_i, AMBER retrieves up to K_w similar earlier episodes. For each association type r, it collects the paths containing these episodes in retrieval order, removes duplicate paths, and keeps at most L candidates in $\Pi_i^{r,\mathrm{cand}}$. This is a semantic candidate filter. The implementation does not maintain a recency registry. We use positive indices for paths in the mathematical description; these indices need not equal the implementation's zero-based identifiers.

For input text x, let $\mathcal{P}_r$ be the fixed regular-expression pool for type r, and p an individual pattern. The indicator $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. Matching is case-insensitive, and each matched pattern contributes once:

$$\sigma^r(x) = \min\left\{1, \frac{1}{3}\sum_{p\in\mathcal{P}_r}\mathbf{1}[p \text{ matches } x]\right\}. \tag{7}$$

The divisor is always three, including for the temporal pool, which contains four patterns. Procedural patterns cover ordering words, step/process terminology, and ordering expressions followed by a next-step cue. Temporal patterns cover temporal connectives, month names, disambiguated references to May, and year/date expressions. Complementary patterns cover additional items, inclusion, and shared collections. Conflicting patterns cover discontinued states, replacements, and explicit corrections. The controller receives both the scores and the matched patterns. The vector $\boldsymbol{\sigma}_i$ in the main text collects these scores for x_i; a controller is invoked for every type even when its signal is zero.

Controller interface and validation. Each type-specific call receives the incoming episode text and timestamp, all four signal summaries, and candidate-path summaries. A summary contains the path identifier, association type, anchor, member count, and member timestamps in path order. Candidate episode texts and images are not included in this controller call. Separate caption fields enter the semantic retrieval payload; they enter this controller only if preprocessing has already included them in the episode text.

The structured output specifies a decision, selected path identifier or new anchor as applicable, confidence, and a reason. The output must provide a valid, nonempty anchor whenever one is required. Malformed outputs, failed calls, or confidence below a construction threshold δ_c produce

$d_i^r = \perp$. The provided dispatcher makes one call per type without automatic retries. A failure for one type does not suppress the episode or decisions for the other types.

Path updates. Let $\Pi_i^r = \{\pi_j \in \Pi_i : r_j = r\}$ denote the paths of type r after processing episode e_i. An episode belongs to at most one path per association type. Anchor identity uses exact equality of its normalized representation after lowercasing text and trimming surrounding whitespace. If a request with $d_i^r = 0$ names an existing anchor of the same type, it is redirected to that path. Let $\bar{d}_i^r$ denote the effective decision after this identity check. Thus, $\bar{d}_i^r = 0$ always denotes creation of a new path, while $\bar{d}_i^r = j$ denotes extension of π_j.

For extension, the episode is added to the stored member sequence. Temporal paths are reordered by parsed timestamps, with ingestion identifiers breaking ties. Records with unparseable timestamps follow valid timestamps and retain ingestion order. Other association types retain ingestion order. Let $(v_1, \ldots, v_m)$ be the resulting sequence of m episodes, with v_ℓ denoting its ℓth episode. Its links are rebuilt as $\mathcal{A}_j^+ = \{(v_\ell, v_{\ell+1})\}_{\ell=1}^{m-1}$, giving $\pi_j^+ = (r_j, a_j, \mathcal{V}_j \cup \{e_i\}, \mathcal{A}_j^+)$. The type-specific update is

$$\Pi_i^r = \begin{cases} (\Pi_{i-1}^r \setminus \{\pi_j\}) \cup \{\pi_j^+\}, & \bar{d}_i^r = j, \\ \Pi_{i-1}^r \cup \{(r, a_i^r, \{e_i\}, \emptyset)\}, & \bar{d}_i^r = 0, \\ \Pi_{i-1}^r, & \bar{d}_i^r = \perp. \end{cases} \tag{8}$$

The updated graph contains $\Pi_i = \bigcup_{r \in \mathcal{R}} \Pi_i^r$ and all observed episodes, including episodes that enter no path.

D.2 Query-Conditioned Activation

Single-mode decision and fallback. The default routing policy uses an LLM activator f_{q} on query text and the same signal pools as construction. It proposes a mode $\tilde{z}_t$, query anchor a_t^q, and confidence $\gamma_t \in [0, 1]$:

$$(\tilde{z}_t, a_t^q, \gamma_t) = f_{\mathrm{q}}\big(x_t^q, \{\sigma^r(x_t^q)\}_{r \in \mathcal{R}}\big). \tag{9}$$

The proposed mode belongs to $\mathcal{R} \cup \{0, \perp\}$. The output includes an anchor when required and a short reason. The activator does not consume query images or retrieved episodes. Images remain available to multimodal semantic retrieval and answer generation.

Let $\nu_t = 1$ indicate a valid structured output, including an admissible mode, finite confidence in $[0, 1]$, and a valid, nonempty anchor for association modes. A proposal of $\perp$ additionally requires an explicit flag that the current input is sufficient. Otherwise, $\nu_t = 0$. With activation threshold δ_z, the accepted mode is

$$z_t = \begin{cases} \tilde{z}_t, & \nu_t = 1 \ \wedge \ \gamma_t \geq \delta_z, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

A failed activator call also yields $z_t = 0$. The supplied code additionally exposes an optional frozen reproduction policy that can override the LLM mode through deterministic query rules. This policy is distinct from the default LLM activation described here.

Anchor agreement and path expansion. For anchors a and a', exact equality of their normalized representations gives $s_a(a, a') = 1$. Otherwise, let $T(a)$ denote the word-token set derived from the normalized anchor representation, retaining its internal token labels. The agreement score is

$$s_a(a, a') = \frac{|T(a) \cap T(a')|}{|T(a) \cup T(a')|}, \tag{11}$$

with score zero when either token set is empty. A seed expands to its stored path only when both the association type and the anchor threshold δ_a match. Expansion retrieves the complete ordered path; seeds reaching the same path are merged into one candidate. The candidate retains the largest seed similarity and its corresponding seed. The supplied implementation has no hop limit or hop-decay factor. With $z_t = 0$, every seed is a singleton candidate; with $z_t = \perp$, retrieval and expansion return no candidates.

D.3 Evidence Chain Ranking

Scores and ordering. Associated candidates use Eq. 4; unmatched singletons use their semantic retrieval score. Candidates are sorted by descending raw score, without length normalization. Equal scores place associated paths before singletons, then use ascending path or episode identifiers. When seeds on one path have equal similarity, the first encountered seed remains the strongest seed.

Budget and truncation. The selector visits this ranked list and accepts a whole candidate if its length fits the remaining episode budget. A later candidate that does not fit is skipped. If the first candidate alone exceeds K, the selector retains a contiguous window of length K around its strongest seed and stops. For a path of length $m > K$, let p be the strongest seed's zero-based position. Among integer starting positions $b \in \{0, \ldots, m-K\}$, the chosen window starts at b^*:

$$b^* = \underset{b \in \{0,\ldots,m-K\}}{\arg\min} \left|2b + K - 1 - 2p\right|, \tag{12}$$

with the smallest b chosen on ties. This chooses the closest centered window, including for even K. Cropping occurs after ranking and retains the candidate score.

The selector subtracts candidate lengths from the remaining budget. Because each episode belongs to at most one path per association type and a query activates one type, the resulting associated candidates are disjoint; unmatched singletons also lie outside those paths. Consequently, this length budget bounds the number of unique episodes. Evidence assembly additionally removes repeated episodes at their first occurrence.

D.4 Association-Aware Evidence Assembly

The organizer follows the selected candidate order and each selected chain's internal episode order. For a multi-episode candidate, it includes the association type, path identifier, anchor, and the episode's position in the original path. It retains episode text, captions, and timestamps, and carries images as multimodal content blocks alongside the serialized text. A singleton is supplied without a multi-episode association descriptor. The organizer itself applies no token-count limit, image-count limit, or image-resizing policy; such controls belong to the downstream generation pipeline.

E Generalization and Robustness

We examine task performance across data sources and answer backbones, then test sensitivity to memory perturbations. The supplementary MemStream analysis uses an aligned evaluation cohort of 779 question trajectories, 1,603 evaluated states, and 824 adjacent transitions for five methods and two answer backbones. These evaluated-state counts are distinct from the dataset's 1,603 annotated information states. The perturbation comparisons cover AMBER and MuRAG.

E.1 Task Generalization across Sources and Backbones

The main results show improved answer quality with both backbones. On MemStream, AMBER's gains over MuRAG, averaged equally across the two backbones, are 3.23% in F1, 3.05% in BLEU-1, 1.19% in EM, and 1.22% in LLM-J (Table 1). Answer continuity also improves: Strict Continuity rises from 15.41% to 19.05% with `gpt-4.1-nano` and from 27.67% to 30.58% with `gpt-5.4-mini`. Thus, the advantage in answering evolving questions persists across the two answer backbones.

Source and scenario variation. The supplementary analysis separates MemStream into its Mem-Gallery and H2HMem sources. Together with the two static benchmarks, this gives four evaluation groups. Across these groups, two backbones, and four answer metrics, AMBER leads in 24 of 32 comparisons (75.00%) and ranks in the top two in 84.40%. Its mean rank is 1.547. Defining each score difference as AMBER minus the strongest baseline in that comparison, the mean difference is $+2.40\%$ and the smallest is -1.33%. Gains are concentrated in the Mem-Gallery-derived streams; performance on the H2HMem-derived streams is closer to the baseline and varies by metric. These results describe performance across sources. The supplementary runs use a frozen configuration for Mem-Gallery and the original H2HMem evaluation pipeline, so they do not establish transfer of a single identical configuration without retuning.

Variation also appears within sources. When scenarios with fewer than 20 evaluated samples are excluded, AMBER's scenario win rates in LLM-J are 31.40% and 40.00% for `gpt-4.1-nano` and `gpt-5.4-mini`, respectively. The corresponding differences, averaged equally across these retained scenarios, are -2.06% and -0.66%. This filtered supplementary analysis uses a different cohort from the main table. Aggregate gains therefore do not imply consistent advantages in every scenario.

Retrieval consistency across backbones. In the frozen supplementary runs, the selected association mode and ordered retrieved episode identifiers agree for all 1,603 matched states (100.00%) across the two backbones. This controls the evidence presented when comparing their answers. It is an observation about these fixed runs, not a guarantee that every memory configuration is independent of its auxiliary model. The MemStream analysis uses one stream seed, 73, and does not measure variability across independently sampled streams.

E.2 Robustness to Memory Perturbations

Duplicate evidence. The duplication experiment evaluates text, visual, and complete-record duplication at four intensities, giving 12 conditions per method. Let m denote the duplication intensity. At the strongest tested intensity, $m = 10$, AMBER outperforms MuRAG by 1.26% in F1, 2.70% in LLM-J, and 2.06% in supporting-evidence recall, averaged across the three duplication types. Supporting-evidence recall measures the fraction of annotated answer support recovered by retrieval. All differences use the reported scores on a 0–100 scale.

The advantage is strongest under heavy duplication. Across all 12 conditions, AMBER has slightly higher F1 and supporting-evidence recall, while MuRAG has slightly higher mean LLM-J. Duplicate pollution remains substantial even for AMBER: at $m = 10$, repeated records still occupy approximately 80% of the ten retrieved positions under text and complete-record duplication. AMBER reduces interference in the strongest tested conditions without eliminating it.

We compare answer scores on the full perturbed evaluation sets. AMBER's matched clean-to-perturbed F1 coverage is only 102/392 for text and complete-record duplication and 70/82 for visual duplication. Comparing performance drops computed from these unequal paired subsets would change the evaluation population. The available logs also do not fully record duplicate-seed merging, so the observed gains cannot be attributed specifically to that operation.

Physical storage order. Eight storage-order permutations cover 18 scenarios and 114 query keys. This perturbation changes physical serialization while preserving episode contents and timestamps. AMBER retains the same selected association mode, ordered final evidence, and answer in all tested permutations (100.00% agreement); MuRAG is also highly stable under this test. This result concerns physical storage order and does not imply robustness to changing the temporal order of the conversation. Results for semantic decoys with high similarity but incorrect facts were incomplete at the analysis cutoff and are excluded from the comparison.

E.3 Label-Preserving and Counterfactual Perturbations

Label-preserving perturbations test whether an answer remains stable when the required information is unchanged. Fact-changing perturbations test whether the answer responds to a revised reference. These are separate requirements: stable output alone is insufficient when the evidence changes.

Label-preserving proxy. The available paired analysis contains 9,708 semantic-interference pairs whose reference labels remain unchanged. Let M_{inv} denote the number of pairs, and let $\hat{a}_i^{\text{clean}}$ and $\hat{a}_i^{\text{pert}}$ be the answers before and after perturbation for pair i. The output-consistency proxy is

$$\text{IC}_{\text{proxy}} = \frac{1}{M_{\text{inv}}} \sum_{i=1}^{M_{\text{inv}}} \text{F1}\big(\hat{a}_i^{\text{clean}}, \hat{a}_i^{\text{pert}}\big), \tag{13}$$

where F1 is the normalized token-overlap score used for answer evaluation. This measures agreement between predictions, so it must be interpreted alongside their correctness against the reference. The paired analysis does not establish a significant answer-invariance advantage for AMBER over

MuRAG. It is a proxy evaluation: the full suite of query paraphrases, evidence paraphrases, irrelevant-episode additions and deletions, equivalent caption rewrites, and physical-position changes has not been completed. We therefore do not report a formal combined metamorphic stability score.

Fact-changing counterfactuals. The counterfactual evaluation contains 400 paired samples across eight perturbation types. For pair i, let a_i^{orig} be the original reference, a_i^{cf} the revised reference, and $\hat{a}_i^{\text{cf}}$ the answer generated from the modified input. With M_{cf} pairs, Counterfactual Flip Accuracy (CFA) is

$$\text{CFA} = \frac{1}{M_{\text{cf}}} \sum_{i=1}^{M_{\text{cf}}} \mathbf{1}\left[\text{F1}(\hat{a}_i^{\text{cf}}, a_i^{\text{cf}}) > \text{F1}(\hat{a}_i^{\text{cf}}, a_i^{\text{orig}})\right], \tag{14}$$

where $\mathbf{1}[\cdot]$ is one when its condition holds and zero otherwise. A tie counts as a failure to flip. CFA tests movement toward the revised reference; it does not require an exactly correct revised answer.

Table 8: Paired counterfactual outcomes on 400 samples. Success is defined by Eq. 14. Entries are counts.

	MuRAG succeeds	MuRAG fails	Total
AMBER succeeds	174	6	180
AMBER fails	12	208	220
Total	186	214	400

AMBER achieves CFA of 45.00%, compared with 46.50% for MuRAG. Only 18 pairs have discordant outcomes: six favor AMBER and twelve favor MuRAG. A two-sided exact McNemar test on these outcomes gives $p = 0.238$, so the observed -1.50% difference does not establish a statistically significant advantage for either method. The complementary, conflicting, and procedural perturbation groups contain only 13, 6, and 16 samples, respectively, limiting type-specific conclusions. The current evidence supports improved performance on evolving question-answering tasks, but does not establish superior overall preservation of metamorphic relations.

F MemStream Annotation and Metrics

F.1 Construction Pipeline

Each question is grounded in its source conversation and final reference answer. The annotation distinguishes information states by the evidence available in the original episode order. Each checkpoint records the current answer and its supporting episodes. Delayed queries remain within the interval for which that answer is valid. Checks enforce increasing prefix lengths, valid evidence identifiers, prefix-grounded answers, and consistency with the source answer at the final state.

F.2 Overall Statistics

The source pool comprises 1,711 Mem-Gallery questions and 2,236 H2HMem questions. The audit identifies 288 and 491 evolution-related questions, respectively, for 779 question trajectories and 1,603 state annotations.

Each retained question receives one category describing how its answer changes as information becomes available: *information introduction*, *answer update*, *conflict transition*, *set accumulation*, *commitment formation*, *event lifecycle*, or *causal closure*. These annotation categories are independent of the four association types used by AMBER. Here, an event is a real-world happening that may be mentioned across several episodes.

F.3 Answer Metric Definitions

Let P be the number of evaluated question trajectories, L_p the number of checkpoints for question p, and $S = \sum_{p=1}^{P} L_p$ the total number of evaluated states. At each state, the generated answer $\hat{a}_{p,\ell}$

is compared with its state-specific reference $a_{p,\ell}$. Write $F_{p,\ell}$ for their token-level F1 and $J_{p,\ell} \in \{0, 0.5, 1\}$ for their reference-guided LLM-J score after the mapping specified in Section 5.1. The process summaries below use checkpoint, question, or transition units as specified, separately from the main table's scenario macro-average. State F1 and State LLM-J are the means $S^{-1}\sum_{p=1}^{P}\sum_{\ell=1}^{L_p} F_{p,\ell}$ and $S^{-1}\sum_{p=1}^{P}\sum_{\ell=1}^{L_p} J_{p,\ell}$. Query-Macro LLM-J first averages each question's states and then averages questions equally: $P^{-1}\sum_{p=1}^{P} L_p^{-1}\sum_{\ell=1}^{L_p} J_{p,\ell}$. None of these averages is an intermediate-only or top-grade success rate.

As in Section 4.3, $b_{p,\ell} = \mathbf{1}[J_{p,\ell} = 1]$ indicates the highest mapped grade, which includes raw grades of 0.75 and 1. *Whole-Trajectory Success* (WTS) requires this grade at every state of a question:

$$\mathrm{WTS} = \frac{1}{P}\sum_{p=1}^{P}\prod_{\ell=1}^{L_p} b_{p,\ell}. \tag{15}$$

For *Hidden Earlier Failure* (HEF), the eligible questions are $\mathcal{H} = \{p : L_p > 1,\ b_{p,L_p} = 1\}$, namely multi-state questions whose final answer receives the top grade. HEF measures the proportion with at least one earlier state below this grade:

$$\mathrm{HEF} = \frac{1}{|\mathcal{H}|}\sum_{p\in\mathcal{H}}\left(1 - \prod_{\ell=1}^{L_p-1} b_{p,\ell}\right). \tag{16}$$

Strict and Soft Continuity weight adjacent transitions equally; Correct Retention and Degradation Rate condition on a top-grade earlier answer. A conditional metric with no eligible examples is undefined and reported as N/A. All answer metrics use the reference valid at each checkpoint, so maintaining the top grade may require changing the answer.

F.4 Evidence Access and Retention

At checkpoint ℓ for question p, $\mathcal{E}_{p,\ell}$ is the annotated supporting episode set and $\widehat{\mathcal{E}}_{p,\ell}$ is the set retrieved under the Top-10 episode budget. For $\ell \geq 2$, define

$$\mathcal{O}_{p,\ell} = \mathcal{E}_{p,\ell-1} \cap \mathcal{E}_{p,\ell}, \qquad \mathcal{U}_{p,\ell} = \mathcal{E}_{p,\ell} \setminus \mathcal{E}_{p,\ell-1}. \tag{17}$$

Here, $\mathcal{O}_{p,\ell}$ contains earlier evidence that remains necessary, and $\mathcal{U}_{p,\ell}$ contains newly required evidence. Newly required does not necessarily mean newly arrived. At an individual checkpoint, support recall uses all of $\mathcal{E}_{p,\ell}$ as its denominator; retained recall and recall of newly required evidence use $\mathcal{O}_{p,\ell}$ and $\mathcal{U}_{p,\ell}$, respectively. Each numerator counts the corresponding evidence found in $\widehat{\mathcal{E}}_{p,\ell}$; a ratio with an empty denominator is undefined. Hit@10 records whether any current support is retrieved, nDCG@10 assesses its rank, and P–R H is the harmonic mean of retained recall and recall of newly required evidence.

Retained recall measures coverage of still-valid earlier evidence. To identify an actual retrieval loss across a transition, an earlier supporting episode must also have appeared in the preceding retrieved set. Answers and retrieved evidence are therefore assessed separately. Outdated evidence is not a retention target unless the current reference still requires it, for example to establish a conflict or replacement. These measures assess evidence access and answer stability across information states. Attributing a failure to association construction additionally requires inspecting graph updates and retrieval decisions.

G Case Studies

We examine four cases to illustrate how cross-episode association helps retrieve evidence missed by MuRAG under the same Top-10 episode budget. Table 9 summarizes support recall, measured as the fraction of annotated supporting episodes retrieved. The boxes show the question, relevant evidence, retrieval decisions, and recorded answers. Episode descriptions and anchor labels are shortened for readability; method names and conversation terminology are normalized to those used in this paper. Unrecorded controller confidence and numerical chain scores are omitted. These are individual retrieval cases, rather than question trajectories from the information-state evaluation.

Table 9: Support recall in the four cases. Counts indicate retrieved supporting episodes divided by all required supporting episodes; recall values are percentages. QA ID identifies the source question–answer pair. Evidence coverage and answer quality are assessed separately.

		MuRAG		AMBER	
Association	QA ID	Count	Recall	Count	Recall
Temporal	7	1/2	50.00	2/2	100.00
Procedural	3	3/9	33.33	9/9	100.00
Complementary	17	1/5	20.00	5/5	100.00
Conflicting	52	0/1	0.00	1/1	100.00

G.1 Temporal Association

The question requires comparing two discussions. Retrieving the later discussion alone leaves the temporal comparison unsupported.

Temporal association: comparing two discussions

Source: `AI_Robotics_Automation_Future_Tech`, QA 7.

Question

Did Julian discuss ethical dilemmas in autonomous driving before or after raising concerns about robots in healthcare? **Reference answer:** After.

Supporting evidence in chronological order

`D2:2` Concerns about robots in healthcare.
↓
`D2:3` Discussion of human involvement in automated decisions.
↓
`D2:4` Ethical dilemmas in autonomous driving.

The annotated support is {`D2:2`, `D2:4`}; `D2:3` records an intervening discussion.

MuRAG: one endpoint is missing

`D2:4` ranks first, and `D2:3` is also retrieved, but `D2:2` is absent from the Top-10 evidence. The missing episode discusses medical robots, human interaction, and diagnostic errors. Support recall is $1/2 = 50.00\%$.

Answer: "Julian discussed ethical dilemmas in autonomous driving before raising concerns about robots in healthcare."

AMBER: retrieve both endpoints through the matched path

The controller selects temporal association with the anchor *human oversight*. Seeds `D2:4` and `D2:3` match the same stored path and yield one candidate evidence chain, with `D2:4` as its strongest seed. Expanding the matched path includes `D2:2`; all three episodes fit within $K = 10$ and are retained in chronological order. Support recall becomes $2/2 = 100.00\%$.

Answer: "Julian discussed ethical dilemmas in autonomous driving after raising concerns about robots in healthcare."

G.2 Procedural Association

Here, a course description is insufficient: the question also asks for the steps performed during the course. The required evidence spans preparation, bread making, and completion.

Procedural association: completing a baking procedure

Source: `Baking_Dessert_Daily_Life_Skill`, QA 3.

Question

What experiences and steps would a person have when attending a baking course?

Supporting evidence in procedural order

D2:1	Enter the classroom.
D2:2	Learn about the tools.
D2:3	Learn about the ingredients.
D2:4	Weigh the flour.
D2:5	Add ingredients, stir, and knead.
D2:6	Ferment the dough.
D2:7	Shape the dough.
D2:8	Bake until golden brown.
D2:9	Take out the finished bread.

MuRAG: course preparation without the subsequent steps

Only D2:1, D2:2, and D2:3 are retrieved, giving support recall of $3/9 = 33.33\%$. The six episodes describing bread making and completion are missing.

Answer: "Attending a baking course involves observing the environment with neatly arranged ovens and tools, learning about various baking tools and molds, understanding baking ingredients and their functions, and gaining hands-on experience in baking techniques. It also includes appreciating the importance of presentation and decoration, which enhances the beauty and emotional appeal of baked goods."

AMBER: retain the ordered procedure

The controller selects procedural association with the anchor *making bread in the baking course*. The three seeds match one stored path and are merged into one candidate evidence chain, whose strongest seed is D2:1. Path expansion includes D2:4 through D2:9. The complete nine-episode chain fits within $K = 10$, giving support recall of $9/9 = 100.00\%$.

Answer: "Go to the baking classroom, learn about the tools and ingredients, weigh the flour, stir and knead the dough, ferment the dough, shape it, and then bake it."

The generated answer now includes weighing, kneading, fermentation, shaping, and baking. It does not explicitly mention taking out the bread in D2:9, despite that episode being retrieved. Thus, complete support recall does not by itself imply that every supported step is stated in the answer.

G.3 Complementary Association

The requested list is distributed across five episodes. Each episode extends the same set of applications, so one relevant application does not substitute for the remaining items.

Complementary association: collecting a complete list

Source: AI_Robotics_Automation_Future_Tech, QA 17.

Question

Please list the applications of AI in cybersecurity as described by Julian.

Supporting evidence under the shared anchor

D4:3	Detect and mitigate threats as they occur.
D4:4	Analyze network traffic and identify anomalies.
D4:7	Automate responses to threats.
D4:9	Identify insider threats.
D4:11	Predict and prevent attacks.

MuRAG: one relevant item without the full set

Only D4:3 is retrieved among the five supporting episodes, giving support recall of $1/5 = 20.00\%$. The other four are missing.

Answer: "AI in cybersecurity is used to detect and mitigate threats in real time, adapt to new attack types through continuous learning, enhance detection of zero-day vulnerabilities, and provide an additional layer of defense alongside human oversight."

This answer omits network-traffic analysis, automated responses, insider-threat detection, and attack prediction from the required list.

AMBER: assemble evidence that extends the same set

The controller selects complementary association with the anchor *applications of AI in cybersecurity*. Seed `D4:3`, ranked third by similarity, matches a stored path containing all five episodes. Expanding this path adds `D4:4`, `D4:7`, `D4:9`, and `D4:11`. The resulting five-episode evidence chain fits within $K = 10$ and is retained in full, giving support recall of $5/5 = 100.00\%$.

Answer: "AI in cybersecurity is used to detect and mitigate threats in real time, analyze network traffic to identify anomalies, predict and prevent attacks, automate responses to threats, and identify insider threats."

The answer covers all five annotated applications. The role of complementary association is to collect distinct contributions about the same anchor; the items need not form a temporal or procedural sequence.

G.4 Conflicting Association Mode

This question asks whether a claim about a final choice contradicts the conversation. Earlier discussion of an option is insufficient to establish that it was ultimately chosen.

Conflicting retrieval mode: checking the final choice

Source: `Astronomy_Physics_Scientific_Experiments_Cosmology`, QA 52.

Question

Please determine whether the following information conflicts with the conversation memory: Leo chose travel faster than light as the natural law to change. **Reference answer:** Yes.

Relevant evidence and the claim being checked

`D11:7`	Faster-than-light travel was considered.
`D11:8`	Faster-than-light travel was explored.
`D11:20`	Infinite energy was finally chosen.

Query claim: Leo finally chose faster-than-light travel.
Decisive supporting evidence: `D11:20` contradicts that claim by identifying a different final choice.

MuRAG: discussion of an option mistaken for a final choice

The Top-10 evidence contains multiple episodes about faster-than-light travel but excludes `D11:20`, which discusses infinite energy and its practicality. Support recall is $0/1 = 0.00\%$.

Answer: "No."

AMBER: include the evidence specifying the final choice

The case record reports that the controller selects the conflicting retrieval mode with the anchor *Leo's chosen natural law*, and that retrieval from seeds including `D11:8` and `D11:7` retains the matched evidence chain containing `D11:20` within $K = 10$. The strongest recorded seed is `D11:8`. Support recall becomes $1/1 = 100.00\%$.

Answer: "Yes."

The contradiction is between the query's claimed final choice and the choice recorded in `D11:20`. Considering faster-than-light travel and later choosing infinite energy can both be true. The case demonstrates retrieval under the selected mode and the need for final-choice evidence; it does not independently validate the semantics of the stored associations.

Across these cases, evidence with high query similarity provides useful seeds but can omit the earlier endpoint of a comparison, subsequent steps of a procedure, additional members of a set, or the

episode that specifies a final choice. With association-routed retrieval, the evidence context includes the missing episodes in these cases. The examples illustrate the resulting differences in evidence access and generated answers; they do not establish aggregate performance or continuity across information states.

H LIMITATIONS

Our evaluation focuses on image-and-text conversations and two answer backbones, using MemStream to study cross-episode evidence retrieval. Although it does not cover all modalities, domains, or intermediate states, this controlled scope isolates the effects of association-based organization. Extending evaluation to audio, video, more domains, and finer-grained states is future work. AMBER's four association types capture common cross-episode dependencies, not an exhaustive taxonomy. Multiple types can coexist, but the current query-conditioned mechanism activates one type to keep evidence paths compact. Semantic candidate retrieval and query-relevance ranking help surface useful evidence; broader association selection and more robust anchor matching remain open improvements. AMBER adds computation for online association construction and activation, as well as storage for organized paths. We control these costs with fixed retrieval budgets and bounded evidence contexts, comparing methods under the same resource settings. Our results assess retrieval and answer quality, not latency or storage advantages. Reducing costs over longer histories will be a next step.

I REPRODUCIBILITY AND AI USE STATEMENT

Reproducibility. We provide the implementation details, prompts, evaluation procedures, and additional experimental results needed to reproduce AMBER in Appendices B–F. Appendix C specifies all LLM prompts used for association construction, query-conditioned activation, evidence assembly, answer generation, and LLM-as-a-Judge evaluation. Appendix D describes the implementation of online association construction, query-conditioned retrieval, evidence-chain ranking, and evidence assembly, including the retrieval budgets and scoring parameters. Appendix F provides the construction procedure, annotation protocol, and metric definitions for MemStream. The main experiments use the publicly available Mem-Gallery and H2HMem benchmarks, with the same input histories and queries provided to all methods. We use fixed model configurations, retrieval budgets, and evaluation procedures across methods, as described in Section 5.1 and Appendix D. The code, configuration files, prompts, and scripts required to reproduce the reported experiments will be provided as supplementary material in an anonymized form during review.

AI Use. Generative AI tools were used to assist with the preparation of this manuscript, including language editing, organization and refinement of scientific writing, and revision of scientific figures. The authors independently reviewed and verified all AI-assisted content, including the technical descriptions, mathematical formulations, experimental procedures, reported results, and references. No AI system was used as an author, and the authors remain fully responsible for the correctness, originality, and integrity of the submitted work. In particular, all experimental results were obtained from the described experimental pipeline and were independently checked by the authors. No AI-generated content was accepted without human verification.