Remembering Through Change: Provenance-Preserving Hypergraph Memory for Multimodal Agents

Anonymous authors
Paper under double-blind review

Abstract

Long-term memory for multimodal agents remains challenging, as it must preserve fine-grained visual evidence while tracking entity evolution across sessions. Existing abstraction-based memories compress images into text, often discarding visually discriminative details, whereas pixel-preserving memories retain raw observations but struggle to reliably associate diverse observations with persistent entity identities and track their evolving states. To address these limitations, we propose **MHG-Mem**, a multimodal hypergraph memory that jointly organizes entity identity, state evolution, and provenance to visual evidence. Specifically, MHG-Mem represents distinct entity mentions and their observed facts as individual nodes, grouping identity-consistent mentions and facts through occurrence hyperedges. It then synthesizes aspect-specific evolution states from these grouped observations and links them to supporting entities, facts, and source evidence through evolutionary and evidential hyperedges. To navigate this structured memory, we introduce hypergraph-based agentic search, which adaptively selects an entry layer, traverses typed hyperedges across memory levels, and inspects source observations when fine-grained verification is required. These designs support both detailed visual verification and reasoning over cross-session state changes. Extensive experiments across three benchmarks demonstrate the effectiveness of MHG-Mem, improving performance by 8.20–9.98% over the strongest baseline.

1 Introduction

Long-term memory enables large language model (LLM) agents to reuse past experience beyond their context windows (Packer et al., 2023; Xu et al., 2025). As agents increasingly browse websites and interact with users through images, multimodal memory becomes essential for understanding user needs and providing personalized assistance (Guo et al., 2026; Liu et al., 2026a). As shown in Fig. 1(a), a web agent encounters several backpacks across sessions, with varying page layouts, product views, prices, and availability. Subsequent queries require *inspecting the main zipper of the purchased backpack* (Q1) or *determining whether an earlier offer for the same backpack was both cheaper and in stock* (Q2). Answering these queries remains challenging. Memory must preserve fine-grained visual evidence while tracking the entity's state across sessions despite variations in its visual observation.

Existing multimodal memories address these requirements only partially. As illustrated in Fig. 1(b), **abstraction-based memory** converts visual observations into textual descriptions or structured records for subsequent consolidation and retrieval (Wang & Chen, 2025; Feng et al., 2026). Such abstractions may omit visual details needed for inspection or for distinguishing similar entities. **Pixel-preserving memory** retains original images, allowing relevant observations to be retrieved and inspected at query time (Lu et al., 2026). Yet the retrieved images remain isolated observations. Nothing records that the product photo in one session and the listing page in another show the same backpack, or how its price and availability evolved. Without reliable associations, an agent may confuse similar entities or incorrectly combine attributes in different observations. These raise the question: *How can*

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

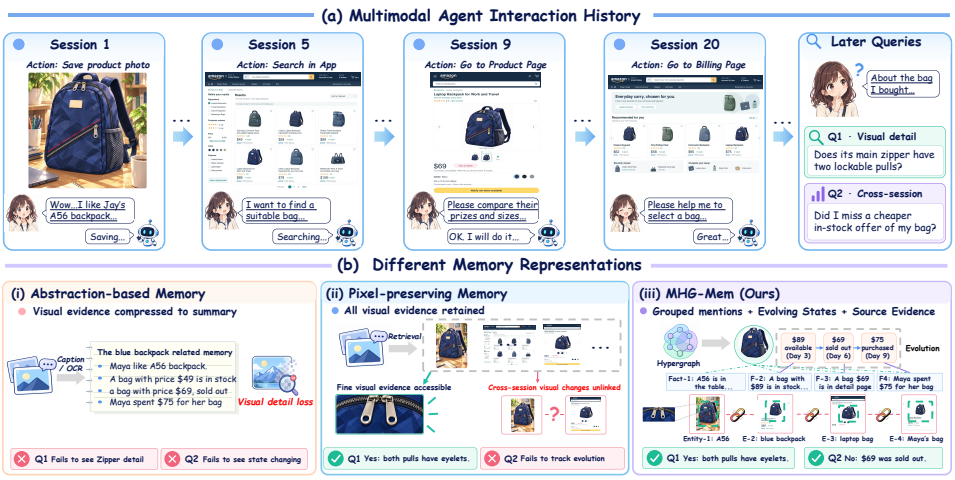

Figure 1: (a) Long-term agent–user multimodal interaction example. (b) Comparison of existing approaches: (i) Abstraction-based memory loses visual details (failing Q1, Q2); (ii) Pixel-preserving memory lacks persistent identity–evolution bindings (failing Q2); (iii) Our MHG-Mem organizes multimodal observations via typed hyperedges, preserving both fine-grained visual crops for inspection and traceable evolution across diverse entity mentions.

multimodal memory organize visually diverse observations of the same entity into coherent state histories while keeping each state traceable to its supporting evidence?

Addressing this question requires preserving individual observations while explicitly organizing how they relate across sessions. The underlying memory structure is inherently *higher-order*: the same entity may be observed differently across sessions, with each observation jointly associating several entity mentions, facts, and visual evidence, while its evolutions must integrate these observations into coherent state histories. These relationships therefore extend beyond isolated entities or facts and naturally form structured groups at different semantic levels. Motivated by this, we propose **MHG-Mem**, a multimodal hypergraph memory that uses typed hyperedges to explicitly organize these higher-order relationships while preserving each observation separately. Specifically, occurrence hyperedges group identity-consistent entity mentions and their associated facts across sessions, evolution nodes summarize aspect-specific state changes from temporally ordered observations, and evolutionary and evidential hyperedges connect each evolution to the participating entity mentions and its supporting facts, respectively. Because these facts retain provenance to the original dialogue and images, each inferred state remains traceable to its source evidence.

This structured memory naturally exposes evidence at multiple granularities, but different queries require different retrieval paths. Q1 ultimately requires inspecting a particular source image, whereas Q2 requires first locating the backpack's state history and then verifying the facts underlying the relevant price and availability states. A fixed retrieval strategy therefore cannot serve both effectively. To this end, we introduce **Hypergraph-based Agentic Search**, which adaptively selects an entry layer, traverses typed hyperedges across entity, evolution, and fact-level memory, and follows provenance links to inspect source observations when necessary. In this way, compact evolution records guide cross-session reasoning, while grounded facts and images provide fine-grained, verifiable evidence. Our contributions are summarized as follows:

- We propose **MHG-Mem**, a multimodal hypergraph memory framework that organizes cross-session observations into coherent entity evolution histories through typed hyperedges, while preserving visual evidence and state-specific provenance.
- We introduce **Hypergraph-based Agentic Search**, which adaptively accesses entity groups, state histories, and their supporting observations to assemble query-relevant evidence across granularities.
- Experiments across three benchmarks and three backbones demonstrate MHG-Mem's effectiveness, improving average performance by 8.20–9.98% over the strongest baseline.

2 Related Work

Long-Term Multimodal Agent Memory. As agents increasingly interact with the physical and digital world, long-term memory research has expanded beyond text-centric systems (Packer et al., 2023; Xu et al., 2025; Kang et al., 2025; Chhikara et al., 2025; Zhong et al., 2024) to multimodal retrieval (Chen et al., 2022; Yeo et al., 2026), episodic video memory (Jeong et al., 2025; Fan et al., 2025), and multimodal conversational memory (Bei et al., 2026; Maharana et al., 2024). Recent benchmarks further examine whether agents can retain action trajectories, and visually evolving states across long-horizon interactions (Liu et al., 2026a; Guo et al., 2026). Existing systems address important aspects of this problem by compressing observations into semantic abstractions (Wang & Chen, 2025; Feng et al., 2026; Jain et al., 2025), organizing experiences around entities or events for cross-session reasoning (Long et al., 2025; Chen et al., 2026; Nguyen et al., 2026; Liu et al., 2026b), and retrieving evidence across modalities or memory stores (Kang et al., 2026; Lu et al., 2026; Tian et al., 2026). Yet these capabilities remain difficult to reconcile: semantic abstraction improves memory organization but may discard visual cues needed to distinguish similar observations, while preserving raw pixels alone does not establish which observations refer to the same entity or how their attributes evolve over time. MHG-Mem jointly addresses these challenges by preserving multimodal occurrences, associating them with persistent entity identities, synthesizing aspect-specific state histories, and grounding each state in supporting facts and source images within a typed hypergraph, enabling agentic search to traverse between compact abstractions and verifiable evidence.

Hypergraph-based Agent Memory. Graph memory has evolved from knowledge graphs (Rasmussen et al., 2025; Gutierrez et al., 2024), toward entity-centered graph such as M3-Agent (Long et al., 2025) and query-aware evidence expansion in GraphMemix (Li et al., 2026). Hypergraph methods then make higher-order membership explicit: HyperGraphRAG (Luo et al., 2025b) represents n-ary facts for corpus-level RAG, HGMEM (Zhou et al., 2025) uses hypergraph working memory for multi-step reasoning, and HyperMem (Yue et al., 2026) organizes long conversations as topic–episode–fact hypergraphs with coarse-to-fine retrieval. Despite their effectiveness, these approaches remain largely confined to text-centric memory organization, limiting their applicability to information-dense and dynamically evolving multimodal agent interactions. In contrast, MHG-Mem uses typed hyperedges to jointly organize persistent multimodal occurrences, entity-specific aspect evolutions and provenance, enabling identity-aware state tracking and fine-grained visual verification within a unified long-term memory framework.

3 Methodology

To track entity states across sessions while preserving their supporting evidence, we propose MHG-Mem, a multimodal hypergraph memory framework. Typed hyperedges organize entity mentions, facts, and state evolutions into reusable histories while retaining observation-specific visual evidence and links from individual states to their sources. Hypergraph-based agentic search then adapts retrieval to the query, navigating these histories and inspecting the underlying evidence as needed. Figure 2 illustrates (a) memory construction, (b) the multimodal hypergraph memory structure, and (c) hypergraph-based agentic search.

3.1 Problem Formulation

Let $\mathcal{D}_{\leq T} = (d_1, \ldots, d_T)$ denote an agent's ordered multimodal interaction history across sessions , where each round $d_t = (x_t, \mathcal{I}_t, \delta_t)$ comprises dialogue text x_t, images $\mathcal{I}_t$, and a recorded date δ_t. A memory construction operator $\mathcal{W}$ maps this history to persistent memory $\mathcal{M}_T = \mathcal{W}(\mathcal{D}_{\leq T})$ before subsequent queries are known. Given a query $q = (x_q, \mathcal{I}_q)$ with optional visual input $\mathcal{I}_q$, a retrieval operator $\mathcal{S}$ selects relevant evidence from $\mathcal{M}_T$. A multimodal agent then generates an answer conditioned on the query and retrieved evidence:

$$E_q = \mathcal{S}(q, \mathcal{M}_T), \qquad \hat{y} \sim p_\theta(\cdot \mid q, E_q), \tag{1}$$

where E_q denotes the retrieved evidence and θ denotes fixed model parameters. The objective is to design $\mathcal{W}$ and $\mathcal{S}$ to support accurate answering of multimodal agent.

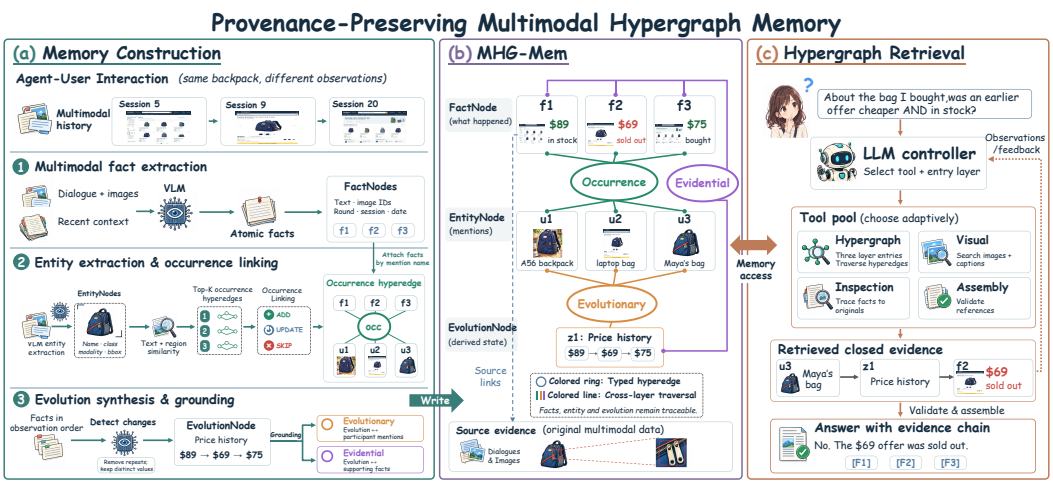

Figure 2: Overview of **MHG-Mem**. (a) Memory construction proceeds through three steps. (b) The three-layer MHG-Mem architecture and three types of hyperedges enable cross-layer traversal and provenance tracing. (c) Hypergraph-based agentic search employs four types of tools to adaptively construct query-aware evidence chains.

3.2 Multimodal Hypergraph Memory Structure

MHG-Mem represents multimodal memory using three node types and three types of n-ary hyperedges. Nodes encode memory units at different levels of granularity, while hyperedges capture cross-level associations and provenance, allowing high-level memory to be traced back to their supporting evidence. We formalize this memory as an attributed hypergraph:

$$\mathcal{H} = (\mathcal{V}, \mathcal{E}, \iota, \tau_V, \tau_E), \quad \mathcal{V} = \mathcal{F} \dot{\cup} \mathcal{U} \dot{\cup} \mathcal{Z}, \quad \mathcal{E} = \mathcal{E}_{\text{occ}} \dot{\cup} \mathcal{E}_{\text{evo}} \dot{\cup} \mathcal{E}_{\text{evd}}, \quad \iota : \mathcal{E} \to 2^{\mathcal{V}} \tag{2}$$

where $\iota(e)$ denotes the set of nodes incident to hyperedge e, while τ_V and τ_E specify the node and hyperedge types, respectively. The contribution of nodes and hyperedges is defined as follows: ❶ `FactNodes` store atomic multimodal observations. ❷ `EntityNodes` store individual textual or visual entity mentions. We denote EntityNodes linked to the same underlying entity as a `referent group` r. ❸ `EvolutionNodes` store state changes in an entity attribute or relation. ❹ `Occurrence hyperedges` bind a referent's EntityNodes to its associated FactNodes. ❺ `Evolutionary hyperedges` connect EvolutionNodes to its associated EntityNodes. ❻ `Evidential hyperedges` connect EvolutionNodes to the supporting FactNodes used for verification.

The complete node, hyperedge, and incidence definitions are given in Appendix B.1. The notation is summarized in Appendix A. This memory structure supports cross-session identity tracking, state synthesis, and provenance tracing while preserving original visual evidence.

3.3 MHG-Mem Construction

Step 1: Multimodal Fact Extraction. In each round d_t, a vision-language model (VLM) extracts atomic multimodal fact descriptions from dialogue and images, using recent within-session context $\mathcal{C}_t$ to resolve references. These descriptions are grounded as FactNodes:

$$\mathcal{F}_t = \text{Ground}_t(\text{VLM}_{\text{fact}}(d_t; \mathcal{C}_t)), \tag{3}$$

where Ground_t creates each $f = (\ell_f, \mathcal{I}_f, t_f, s_f, \delta_f)$ by attaching the current round's image identifiers and source metadata to the extracted description ℓ_f.

Step 2: Multimodal Entity Extraction and Occurrence Linking. A VLM extracts multimodal entity mentions from d_t. For each mention u, we score an occurrence hyperedge by independently maximizing textual and visual similarity over its constituent EntityNodes:

$$s(u, e_r^{\text{occ}}) = w_T \max_{v \in \mathcal{U}_r} c_T(u, v) + w_I \max_{v \in \mathcal{U}_r^{\text{vis}}} c_I(u, v). \tag{4}$$

Here, c_T and c_I are cosine similarities between mention-name and image-region embeddings, respectively, and $\mathcal{U}_r^{\text{vis}}$ contains the visual mentions in $\mathcal{U}_r$. Conditioned on d_t, u, and the top-K hyperedges ranked by s, the model selects one of three actions for occurrence linking: `UPDATE` adds u to the mention set $\mathcal{U}_r$ of a matching referent group, `ADD` creates a new group r with $\mathcal{U}_r = \{u\}$ and its occurrence hyperedge, and `SKIP` rejects an invalid entity mention. Each accepted mention remains a distinct EntityNode, retaining its visual region when available. We then attach facts in $\mathcal{F}_t$ to occurrence hyperedges by mention-name matching:

$$\mathcal{F}_r \leftarrow \mathcal{F}_r \cup \{f \in \mathcal{F}_t : A(f, e_r^{\text{occ}}) = 1\}. \tag{5}$$

Here, $A(f, e_r^{\text{occ}}) \in \{0, 1\}$ is a lexical matching indicator that equals 1 iff ℓ_f contains the name of at least one EntityNode in $\mathcal{U}_r$. The resulting membership is $\iota(e_r^{\text{occ}}) = \mathcal{U}_r \cup \mathcal{F}_r$.

Step 3: Evolution Synthesis and Grounding. To track the evolving states of the same entity across sessions, we synthesize the facts grouped by each occurrence hyperedge into state histories with explicit evidence grounding. For each occurrence hyperedge e_r^{occ}, candidate EvolutionNodes $\widehat{\mathcal{Z}}_r$ are generated from the associated facts $\mathcal{F}_r$:

$$\widehat{\mathcal{Z}}_r = \text{LLM}_{\text{evo}}(\text{Order}(\mathcal{F}_r)), \tag{6}$$

where Order follows observation-order t_f in $\mathcal{D}_{\leq T}$. The model identifies changes in the same attribute or relation a_z across sessions, excluding repeated descriptions and incidental scene variation. Each retained $z \in \widehat{\mathcal{Z}}_r$ records the subject referent $r_z = r$, summary σ_z, update type κ_z (replacement, update, or addition), and state sequence

$$S_z = \big((v_j, f_j)\big)_{j=1}^{m_z}, \qquad \mathcal{F}_z = \{f_j : 1 \leq j \leq m_z\} \subseteq \mathcal{F}_{r_z}, \tag{7}$$

where v_j is the changing value extracted for a_z, and the m_z states follow observation order. Retention requires valid references within $\mathcal{F}_{r_z}$ and at least two distinct values.

The evolutionary hyperedge e_z^{evo} then connects z to subject mentions $\mathcal{U}_z^{\text{subj}} \subseteq \mathcal{U}_{r_z}$ and coactor mentions $\mathcal{U}_z^{\text{co}}$. Coactors come from cited rounds of other groups $r' \neq r_z$ sharing supporting facts, $\mathcal{F}_{r'} \cap \mathcal{F}_z \neq \varnothing$. The evidential hyperedge e_z^{evd} links z to its supporting facts $\mathcal{F}_z$, with S_z retaining the individual state–fact correspondence. Please refer to Appendix B.2 for more details of MHG-Mem construction.

3.4 Hypergraph-based Agentic Search

Since different queries require integrating evidence at different levels of granularity, fixed retrieval and graph traversal still struggle to get the optimal evidence chain (Luo et al., 2025a). We therefore implement $\mathcal{S}$ as agentic search that adaptively selects a memory layer, traverses hyperedges across memory layers, and follows provenance links to sources.

Adaptive Search. Given a multimodal query q with optional images and search history $h_k = ((a_i, o_i))_{i<k}$, an LLM controller π_θ selects a tool call a_k and receives observation o_k:

$$a_k \sim \pi_\theta(\cdot \mid q, h_k), \qquad o_k = \text{Execute}(a_k, \mathcal{M}_T). \tag{8}$$

Four tool types support this agent loop: (i) Hypergraph Tool (ii) Visual Tool (iii) Inspection Tool and (iv) Assembly Tool. Table 4 shows the tool details, the search process is outlined in Algorithm 2.

Hypergraph Retrieval. Each memory layer provides an isolated search tool (as described in Table 4), with hyperedge traversal among retrieved entities, evolutions and facts.

Fact-layer Search. For a search query w generated by the LLM, we rank $f \in \mathcal{F}$ using

$$s_F(w, q, f) = \lambda_T s_T(w, f) + \lambda_{TI} s_{TI}(w, f) + \lambda_{II} s_{II}(q, f), \tag{9}$$

where s_T, s_{TI}, and s_{II} measure text–text, text–image, and image–image cosine similarity with weights λ_T, λ_{TI}, and λ_{II}. Visual terms take the best match over associated images, including $\mathcal{I}_q$ for s_{II}. Retrieved facts retain their source references for inspection.

Entity-layer Search. A name or query image identifies referent groups through maximum textual similarity to referent entity names and visual similarity over $\mathcal{U}_r^{\text{vis}}$ (equation 4). We

then traverse each matched occurrence hyperedge e_r^{occ} from the entity layer to the fact layer $\mathcal{F}_r = \iota(e_r^{\text{occ}}) \cap \mathcal{F}$. This gathers observations of the same underlying entity across rounds.

Evolution-layer Search. To retrieve how an entity changes over time, we first identify its referent group r from a name or query image and retrieve the EvolutionNodes belonging to that group ($r_z = r$). The evolutionary hyperedge e_z^{evo} connects each such node to mentions of its subject entity. We then compare the query text x_q with each node's attribute a_z and summary σ_z to assess relevance:

$$s_Z(q, z) = \cos\big(\phi_T(x_q), \phi_T(a_z \parallel \sigma_z)\big). \tag{10}$$

Here, ϕ_T encodes text, $\cos(\cdot)$ denotes cosine similarity, and $\parallel$ concatenates the attribute and summary. For each selected z, we traverse its evidential hyperedge e_z^{evd} from the evolution layer to the fact layer, getting the supporting facts $\mathcal{F}_z$ in the order recorded by S_z.

Cross-layer Traversal. A traversal step moves from a retrieved node $v \in \mathcal{V}$ to another memory layer through incident hyperedges. Given selected hyperedges $\mathcal{E}' \subseteq \mathcal{E}$ and a target layer $\mathcal{V}' \in \{\mathcal{F}, \mathcal{U}, \mathcal{Z}\}$ with $v \notin \mathcal{V}'$, the cross-layer traversal can be formalized as:

$$\text{Traverse}(v, \mathcal{E}', \mathcal{V}') = \bigcup\nolimits_{e \in \mathcal{E}',\, v \in \iota(e)} (\iota(e) \cap \mathcal{V}'). \tag{11}$$

Source Inspection and Evidence Assembly. The Inspection Tool follows $\rho(\cdot)$ from retrieved FactNodes to the original observations, enabling the controller to inspect fine-grained evidence for queries that require detailed information or visual verification. For an EvolutionNode z, traversing e_z^{evd} to supporting facts $f \in \mathcal{F}_z$ and then following $\rho(f)$ keeps its state history traceable to the original observations. The Assembly Tool organizes selected evidence and image references, together with optional summary, into an evidence chain E_q. After reference validation, this chain is submitted to the QA model with the cited textual and visual evidence to answer q as in Equation equation 1.

4 Experiment

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate on three multimodal memory benchmarks: MemEye (Guo et al., 2026), Mem-Gallery (Bei et al., 2026), and WorldMemArena-S (subset) (Liu et al., 2026a). We report exact match (EM) for MemEye's multiple-choice (MCQ) questions, token-level F1 for the other two benchmarks, and BLEU-1 (B-1) and LLM-as-a-judge (Judge) score for open-ended answers on all three. For memory quality evaluation, we follow (Liu et al., 2026a) using the same judging rubrics, recall, and retrieval coverage. Dataset and metric details are provided in Appendix C.1.

Implementation Details. We use Qwen3-VL-8B (Bai et al., 2025), GPT-4.1-mini (OpenAI, 2025), and Qwen3.8-27B (Qwen Team, 2026) as different memory backbones, GPT-5.4-mini (OpenAI, 2026) as LLM-as-a-judge model. MHG-Mem uses all-MiniLM-L6-v2 for text embeddings and SigLIP2 (Tschannen et al., 2025) for visual and cross-modal embeddings. For a fair comparison, we reproduce all baseline methods with the same model setting and running environment of MHG-Mem. Results are averaged over 3 runs. Additional implementation settings are provided in Appendix C.2.

Baselines. We compare with four unimodal memory methods using dense image caption: A-Mem (Xu et al., 2025), MemoryOS (Kang et al., 2025), SimpleMem (Liu et al., 2026c), and HyperMem (Yue et al., 2026); and four multimodal memory or RAG methods: M2A (Feng et al., 2026), MMA (Lu et al., 2026), SRAG (Guo et al., 2026), and UniversalRAG (Yeo et al., 2026). Detailed baseline descriptions are deferred to Appendix C.3.

4.2 Main Results

Table 1 compares MHG-Mem with eight memory systems across three benchmarks and three backbones. Figure 3 shows memory quality on WorldMemArena-S and fine-grained task performance on WorldMemArena-S and Mem-Gallery. We highlight four key findings:

Table 1: **Comparison with state-of-the-art memory systems across different backbone models.** We report Exact Match (**EM**), BLEU-1 (**B-1**), and LLM-Judge (**Judge**) on **MemEye**, **Mem-Gallery**, and **WorldMemArena-S**. **Avg. J.** denotes the average Judge score across the three benchmarks. Bold and underlined results indicate the best and second-best performance, respectively.

Type	Method	Venue	MemEye			Mem-Gallery			WorldMemArena-S			Avg. J.
			EM	B-1	Judge	F1	B-1	Judge	F1	B-1	Judge	
Qwen3-VL-8B												
Unimodal	A-Mem	NIPS'25	45.15	19.22	29.38	47.68	41.76	66.93	**25.23**	12.16	46.57	47.63
	MemoryOS	EMNLP'25	40.23	17.22	26.28	39.36	35.28	58.86	10.24	3.50	20.18	35.11
	SimpleMem	ICML'26	44.27	16.04	31.26	37.36	33.12	52.43	14.33	4.71	33.50	39.06
	HyperMem	ACL'26	42.31	16.84	28.42	41.05	32.80	62.56	20.06	12.33	40.45	43.81
Multimodal	M2A	Arxiv'26	39.96	10.40	18.41	18.19	14.03	37.65	23.02	11.31	41.64	32.57
	MMA	AACL'26	46.92	20.62	38.17	34.80	30.34	57.49	21.73	10.82	38.95	44.87
	SRAG	Arxiv'26	56.78	24.87	41.80	40.74	35.89	63.62	23.53	11.21	43.93	49.78
	UniversalRAG	ACL'26	57.48	23.92	42.09	51.05	46.39	66.23	14.96	5.06	36.33	48.21
Ours	**MHG-Mem**	–	**61.26**	**26.52**	**46.28**	**66.00**	**60.90**	**76.89**	23.17	**15.29**	**50.78**	**57.98** (+8.20)
GPT-4.1-mini												
Unimodal	A-Mem	NIPS'25	49.73	21.11	38.27	44.08	38.37	64.35	33.39	17.54	60.40	54.34
	MemoryOS	EMNLP'25	45.75	13.12	34.91	43.09	37.76	64.64	17.46	6.63	35.87	45.14
	SimpleMem	ICML'26	50.34	9.42	42.59	34.73	28.90	58.33	23.75	14.71	56.13	52.35
	HyperMem	ACL'26	47.82	19.64	36.43	44.57	33.42	66.76	24.91	15.62	52.04	51.74
Multimodal	M2A	Arxiv'26	52.31	6.31	44.50	15.04	10.28	51.81	29.26	18.55	53.38	49.90
	MMA	AACL'26	50.96	16.32	43.81	26.99	20.60	60.87	30.10	16.33	51.15	51.95
	SRAG	Arxiv'26	58.36	**31.48**	50.54	41.07	35.03	67.06	30.47	18.34	54.27	57.29
	UniversalRAG	ACL'26	60.49	11.23	52.61	48.36	41.90	70.63	26.44	13.03	50.95	58.06
Ours	**MHG-Mem**	–	**66.84**	22.36	**59.73**	**52.58**	**45.75**	**81.56**	**33.61**	**21.52**	**62.57**	**67.95** (+9.89)
Qwen3.8-27B												
Unimodal	A-Mem	NIPS'25	45.10	18.13	33.64	60.56	54.37	75.98	27.99	13.92	57.49	55.71
	MemoryOS	EMNLP'25	42.24	16.67	28.25	42.82	37.49	70.19	10.17	3.76	20.95	39.80
	SimpleMem	ICML'26	47.84	5.37	37.40	49.43	43.26	68.70	18.95	7.34	46.33	50.81
	HyperMem	ACL'26	50.56	21.02	39.66	47.23	36.64	73.03	22.84	10.61	50.57	54.42
Multimodal	M2A	Arxiv'26	48.23	19.51	35.90	20.46	15.49	65.30	28.15	14.14	55.58	52.25
	MMA	AACL'26	56.47	24.95	47.30	35.18	29.03	69.32	26.66	13.35	52.80	56.47
	SRAG	Arxiv'26	66.95	28.26	53.37	41.45	35.68	75.80	28.78	13.97	59.69	62.95
	UniversalRAG	ACL'26	67.85	13.45	60.64	57.30	50.51	76.18	22.11	9.34	53.07	63.30
Ours	**MHG-Mem**	–	**78.76**	**39.54**	**65.93**	**66.07**	**59.84**	**87.85**	**30.48**	**21.21**	**66.05**	**73.28** (+9.98)

(1) MHG-Mem consistently outperforms existing memory systems across benchmarks and backbones. MHG-Mem achieves the highest Judge score in every benchmark–backbone setting, with Avg. J. scores of 57.98% (**+8.20%**), 67.95% (**+9.89%**), and 73.28% (**+9.98%**) on Qwen3-VL-8B, GPT-4.1-mini, and Qwen3.8-27B, respectively. Moreover, MHG-Mem achieves the best or second-best results on metrics including EM, BLEU-1, and F1. These consistent improvements across models and benchmarks demonstrate the effectiveness and generalization of provenance-preserving multimodal hypergraph memory.

(2) Effective visual memory requires structured access to preserved observations. On the visual-centric MemEye benchmark, MHG-Mem with Qwen3.8-27B achieves 78.76% EM, compared with 67.85% for UniversalRAG and 66.95% for SRAG. A similar advantage appears in visually demanding categories in Figure 3(b): MHG-Mem achieves 82% versus 60% on visual reasoning (VR) in Mem-Gallery and 87% versus 83% on visual fact recall (VFR) in WorldMemArena-S. These results suggest that retaining images alone is insufficient when relevant observations cannot be reliably associated with the underlying entity. Occurrence hyperedges address this limitation by grouping visually diverse mentions of the same referent while preserving direct access to observation-specific visual evidence.

(3) Substantial performance gains on temporal and cross-modal reasoning. MHG-Mem achieves the best performance in 17 of the 19 task categories in Figure 3(b), with particularly large gains on tasks requiring integration of observations across sessions. On WorldMemArena-S, temporal reasoning (TR) improves from 48% to 87% over the strongest baseline, while dynamic update (DU), test-time learning (TTL), and cross-modal reasoning (CMR) improve by 17%, 17%, and 15%, respectively. On Mem-Gallery, visual reasoning

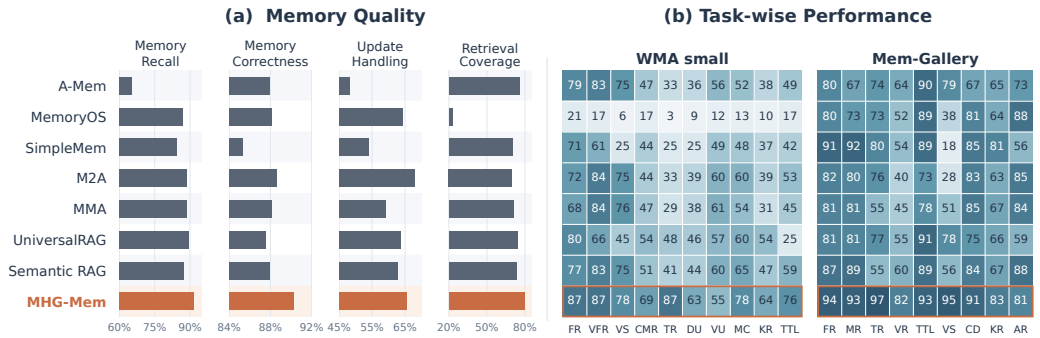

Figure 3: (a) Memory quality and (b) fine-grained tasks performance on WorldMemArena-S (WMA-S) and Mem-Gallery with Qwen3.8-27B.

(VR) and temporal reasoning (TR) improve by 18% and 17%. These gains are supported by our evolution design, which summarize time-ordered entity states to guide retrieval, together with evidential hyperedges that preserve state–fact correspondence and enable provenance tracing from synthesized states back to the originating dialogue and images.

(4) Memory quality and evidence accessibility improve together. As shown in Figure 3(a), MHG-Mem achieves the highest memory recall, memory correctness, and retrieval coverage among all compared systems. The simultaneous improvement in memory quality and coverage indicates that MHG-Mem preserves useful information accurately while keeping its supporting evidence readily accessible.

4.3 Ablation Study

Tables 2 and 3 evaluate memory construction and retrieval components using the average LLM-judge score across the three benchmarks. Memory components are introduced cumulatively, whereas retrieval components are removed individually. Specifically, **w/o Agentic Search** replaces agentic search with three-layer top-k retrieval; **w/o Inspection Tool**, **w/o Cross-layer Traversal**, and **w/o Visual Tool** disable the corresponding functionality.

Memory-side ablation. **(1) Effectiveness of the entity node with occurrence hyperedge.** The entity layer with occurrence hyperedges improves performance by 5.48% over fact-only memory. Occurrence hyperedges account for 4.78% of this gain, with improvements across all three benchmarks. These hyperedges connect mentions and facts associated with the same entity, enabling retrieval of its observations across sessions. **(2) Effectiveness of the evolution node with evolutionary hyperedge.** The evolution layer with evolutionary hyperedges yields a further 3.37% improvement. This representation makes state histories directly retrievable and associates them with entities in subject and coactor roles, supporting reasoning about changes across sessions. **(3) Effectiveness of the evidential hyperedge.** Evidential hyperedges provide an additional 1.56% gain by linking evolution states to their supporting facts. These connections preserve provenance and allow retrieved states to be checked against the observations from which they were derived.

Retrieval-side ablation. **(1) Effectiveness of agentic search.** Three-layer top-k retrieval underperforms agentic search by 5.10%, the largest decline among the retrieval ablations. This result supports query-adaptive selection of retrieval entry points and subsequent evidence expansion. **(2) Effectiveness of the inspection tool.** Without source inspection, performance falls by 3.44%, the largest decline among individual tool ablations. Access to original dialogue and images enables the agent to verify fine-grained evidence that facts and evolution summaries may omit. **(3) Effectiveness of cross-layer traversal.** Removing cross-layer traversal reduces performance by 3.09%. Traversal connects entities to their associated facts and evolution records to supporting observations, allowing the agent to assemble evidence across memory granularities. **(4) Effectiveness of the visual tool.** Performance decreases by 1.37% when image retrieval is disabled. Direct visual search pro-

Table 2: Ablation of memory side with Qwen3.8-27B.

Layers	Hyperedges	MemEye	Mem-Gallery	WMA-S	Avg.
$\mathcal{F}$		53.81	76.08	58.72	62.87
$\mathcal{F} \cup \mathcal{U}$		54.06	77.54	59.10	63.57
$\mathcal{F} \cup \mathcal{U}$	$\mathcal{E}_{\text{occ}}$	60.69	82.13	62.24	68.35
$\mathcal{F} \cup \mathcal{U} \cup \mathcal{Z}$	$\mathcal{E}_{\text{occ}} \cup \mathcal{E}_{\text{evo}}$	64.70	85.48	64.97	71.72
$\mathcal{F} \cup \mathcal{U} \cup \mathcal{Z}$	$\mathcal{E}_{\text{occ}} \cup \mathcal{E}_{\text{evo}} \cup \mathcal{E}_{\text{evd}}$	**65.93**	**87.85**	**66.05**	**73.28**

Table 3: Ablation of retrieval side with Qwen3.8-27B.

Method	MemEye	Mem-Gallery	WMA-S	Avg.
w/o Agentic Search	60.59	82.06	61.90	68.18
w/o Inspection Tool	61.77	84.64	63.12	69.84
w/o Cross-layer Traversal	62.33	84.49	63.74	70.19
w/o Visual Tool	63.52	86.98	65.23	71.91
MHG-Mem (Full)	**65.93**	**87.85**	**66.05**	**73.28**

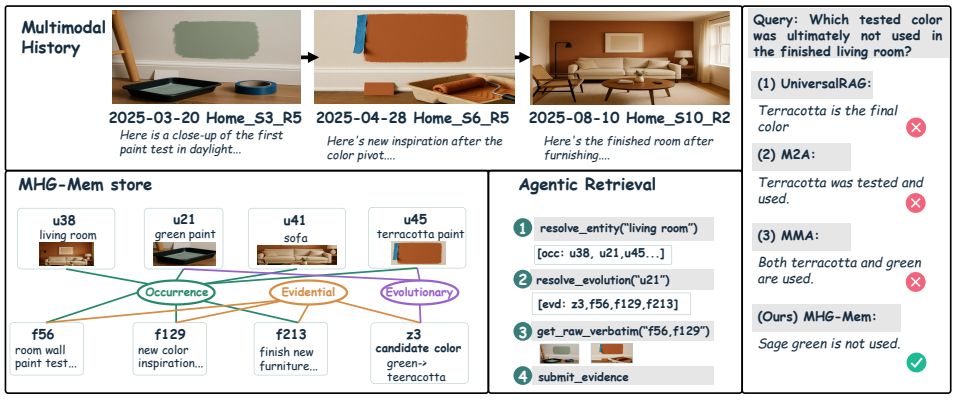

Figure 4: Case study in a home renovation scenario. MHG-Mem tracks wall paint color changes through evolution nodes and traces supporting facts to source images via evidential hyperedges, forming a verifiable evidence chain.

vides a complementary route to relevant observations, extending evidence access beyond entity- and evolution-centred retrieval.

4.4 Case Analysis

Figure 4 illustrates the advantages of MHG-Mem in a home renovation scenario involving successive living room color schemes. To identify the scheme the user did not adopt, the memory system must trace the progression from color trials through revisions to completion and inspect the corresponding images. UniversalRAG and M2A retrieve bedroom images by similarity but miss the initial color information, while MMA preserves the pixels without tracking the evolving schemes. MHG-Mem identifies `paint` as the tracked entity, preserves its mentions across rounds, and links atomic facts through occurrence hyperedges. Evolution nodes record the change in candidate color from sage green to terracotta. During retrieval, the agent follows evidential hyperedges to `f56` and `f129`, then inspects their source images to verify the colors and answer correctly. **Please refer to Appendix D for more experimental results and analysis.**

Conclusion

We presented MHG-Mem, a multimodal hypergraph memory framework that organizes visually diverse observations of the same entity into coherent state histories while preserving fine-grained visual evidence. Typed n-ary hyperedges connect entity mentions, atomic facts, and evolution records, keeping individual states traceable to their supporting observations. Hypergraph-based agentic search adaptively navigates these memory levels, using state histories to guide retrieval and original dialogue and images to verify evidence. Ablations demonstrate the contributions of entity associations, explicit state histories, evidence grounding, and adaptive retrieval. These findings support the joint design of memory organization and evidence access for reasoning over evolving multimodal experience.

AI use statement

In this work, we used generative AI tools solely for language polishing, grammatical correction, and limited code refinement. We did not use generative AI tools for research idea

generation and methodology design. Any AI-assisted code modifications were restricted to implementation-level improvements, such as code restructuring and debugging support, without changing the underlying methodology. All AI-assisted revisions were manually reviewed and verified by the authors to ensure technical accuracy and consistency with the original content.

Reproducibility statement

The main results are reproducible using the code provided in our anonymous repository: `https://anonymous.4open.science/r/MHG_Mem-26C8`. Detailed descriptions of the implementation, such as model settings, hyperparameters, evaluation, and prompt design, are provided in Appendix C.2, C.3, and E. The repository also contains the complete implementation, evaluation scripts, and instructions.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631, 2025. URL `https://arxiv.org/abs/2511.21631`.

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. arXiv preprint arXiv:2601.03515, 2026. URL `https://arxiv.org/abs/2601.03515`.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. 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/`.

Yijun Chen, Yaqi Zheng, Yanya Li, Boyi Xiao, Buqiang Xu, Shuofei Qiao, Jizhan Fang, Xinle Deng, Yunzhi Yao, Xuehai Wang, Liuxin Zhang, Hui Li, Huajun Chen, and Shumin Deng. EM^{2}Mem: Event-Centric Multimodal Memory for Large Language Models. arXiv preprint arXiv:2609.00551, 2026. URL `https://arxiv.org/abs/2609.00551`.

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.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A Memory-Augmented Multimodal Agent for Video Understanding. In *Computer Vision – ECCV 2024*, volume 15080 of *Lecture Notes in Computer Science*, pp. 75–92, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-031-72670-5_5. URL `https://doi.org/10.1007/978-3-031-72670-5_5`.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, and Wentao Zhang. M2A: Multimodal Memory Agent with Dual-Layer Hybrid Memory for Long-Term Personalized Interactions. arXiv preprint arXiv:2602.07624, 2026. URL `https://arxiv.org/abs/2602.07624`.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, and Ruixiang Tang. MemEye: A Visual-Centric Evaluation Framework for Multimodal Agent Memory. arXiv preprint arXiv:2605.15128, 2026. URL https://arxiv.org/abs/2605.15128.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hkujvAPVsg.

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.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-Augmented Generation over Video Corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1096. URL https://aclanthology.org/2025.findings-acl.1096/.

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.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI Agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Geng Li, Yuhao Wang, Dong Li, Jianye Hao, and Yuxin Peng. GraphMemix: Query-Aware Evidence Forests for Long-Term Multimodal Agent Memory. arXiv preprint arXiv:2608.26983, 2026. URL https://arxiv.org/abs/2608.26983.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. WorldMemArena: Evaluating Multimodal Agent Memory Through Action-World Interaction. arXiv preprint arXiv:2605.29341, 2026a. URL https://arxiv.org/abs/2605.29341.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, Mingyu Ding, and Huaxiu Yao. Omni-SimpleMem: Autoresearch-Guided Discovery of Lifelong Multimodal Agent Memory. arXiv preprint arXiv:2604.01007, 2026b. URL https://arxiv.org/abs/2604.01007.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient Lifelong Memory for LLM Agents. arXiv preprint arXiv:2601.02553, 2026c. URL https://arxiv.org/abs/2601.02553.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory. arXiv preprint arXiv:2508.09736, 2025. URL https://arxiv.org/abs/2508.09736.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. MMA: Multimodal Memory Agent. arXiv preprint arXiv:2602.16493, 2026. URL https://arxiv.org/abs/2602.16493.

Haoran Luo, Haihong E, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, and Anh Tuan Luu. Graph-R1: Towards Agentic GraphRAG Framework via End-to-end Reinforcement Learning. arXiv preprint arXiv:2507.21892, 2025a. URL https://arxiv.org/abs/2507.21892.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 152206–152234. Curran Associates, Inc., 2025b. doi: 10.52202/085713-5089. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/df55ee6e59f8ac4a625219e11fe9ddba-Paper-Conference.pdf.

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 (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Viet Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal Visual Memory from Explicit and Implicit Evidence. arXiv preprint arXiv:2605.28806, 2026. URL https://arxiv.org/abs/2605.28806.

OpenAI. GPT-4.1 Mini. OpenAI API model documentation, 2025. URL https://developers.openai.com/api/docs/models/gpt-4.1-mini.

OpenAI. GPT-5.4 Mini. OpenAI API model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini.

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

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Preston 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.

Zhoujin Tian, Hao Zhang, Yao Tian, Cheng Chen, Yakun Li, Lei Zhang, and Xiaofang Zhou. Beyond Retrieval: Analytic Memory for Multimodal Agents. arXiv preprint arXiv:2607.29440, 2026. URL https://arxiv.org/abs/2607.29440.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv preprint arXiv:2502.14786, 2025. URL https://arxiv.org/abs/2502.14786.

Yu Wang and Xi Chen. MIRIX: Multi-Agent Memory System for LLM-Based Agents. arXiv preprint arXiv:2507.07957, 2025. URL https://arxiv.org/abs/2507.07957.

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, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.177. URL https://aclanthology.org/2026.acl-long.177/.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. HyperMem: Hypergraph Memory for Long-Term Conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35237–35254, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1627. URL https://aclanthology.org/2026.acl-long.1627/.

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.

Chulun Zhou, Chunkang Zhang, Guoxin Yu, Fandong Meng, Jie Zhou, Wai Lam, and Mo Yu. HGMEM: Hypergraph-based Working Memory to Improve Multi-step RAG for Long-Context Complex Relational Modeling. arXiv preprint arXiv:2512.23959, 2025. URL https://arxiv.org/abs/2512.23959.

Appendix Contents

A Notation Definition **15**

B Additional Method Details **16**

B.1 Detailed Hypergraph Memory Schema 16

B.2 Detailed Memory Construction 16

B.3 Agentic Search and Tool Specification 18

B.4 Provenance Preservation 18

C Additional Experimental Details **19**

C.1 Benchmarks and Metrics 19

C.2 Implementation Details 21

C.3 Baselines 22

D Additional Experiment Results and Analysis **23**

D.1 Retrieval Tool Analysis 23

D.2 Cost Analysis 24

D.3 Memory Quality and Fine-grained Analysis on More Backbones 25

D.4 Hyperparameter Analysis 25

D.5 More Case Studies 26

E Prompt Templates **28**

E.1 Atomic fact construction 28

E.2 Multimodal entity extraction 29

E.3 Occurrence linking 29

E.4 State evolution synthesis 30

E.5 Hypergraph-based agentic search 31

A Notation Definition

Symbol	Meaning	*Symbol*	Meaning
$\mathcal{D}_{\leq T}$	Ordered multimodal history	d_t	Interaction round t
x_t	Dialogue text in round t	$\mathcal{I}_t$	Images in round t
δ_t	Date of round t	$\mathcal{C}_t$	Recent within-session context
$q = (x_q, \mathcal{I}_q)$	Multimodal query	$\mathcal{W}, \mathcal{S}$	Memory construction and retrieval operators
$E_q, \hat{y}$	Retrieved evidence and generated answer	θ	Fixed model parameters
$\mathcal{H}$	Attributed multimodal hypergraph	$\mathcal{M}_T$	Persistent memory at history length T
$\mathcal{V}, \mathcal{E}$	Node and hyperedge sets	ι	Hyperedge incidence map
τ_V, τ_E	Node-type and hyperedge-type maps	ρ	Source-round reference map
$\mathcal{F}$,$\mathcal{U}, \mathcal{Z}$	FactNode, EntityNode, and EvolutionNode sets	$\mathcal{K}$	Stored embedding indices
$\mathcal{E}_{\text{occ}}, \mathcal{E}_{\text{evo}}, \mathcal{E}_{\text{evd}}$	Occurrence, evolutionary, and evidential hyperedge families	r	Referent group
$f = (\ell_f, \mathcal{I}_f, t_f, s_f, \delta_f)$	FactNode and its text, images, round, session, and date	$u = (n_u, c_u, \mu_u, i_u, b_u)$	EntityNode and its mention attributes
$\mathcal{U}_r, \mathcal{F}_r$	Mentions and facts associated with referent r	e_r^{occ}	Occurrence hyperedge of r
$z = (r_z, a_z, \sigma_z, S_z, \kappa_z)$	EvolutionNode and its subject, aspect, summary, state trace, and update type	$\mathcal{F}_z$	Facts supporting evolution z
$S_z = ((v_j, f_j))_{j=1}^{m_z}$	Ordered state–fact trace of z	v_j, m_z	State value at position j and number of recorded states
$\mathcal{U}_z^{\text{subj}}$,$\mathcal{U}_z^{\text{co}}$	Subject and coactor mentions of z	$e_z^{\text{evo}}, e_z^{\text{evd}}$	Evolutionary and evidential hyperedges
a_k, o_k, h_k	Tool action, observation, and search history at step k	π_θ	LLM search controller
$B_{\text{step}}, B_{\text{img}}$	Search-step and image budgets	w	Controller-generated fact-search string
$\mathcal{F}_t, \widehat{\mathcal{Z}}_r$	Current-round facts and candidate evolutions	A	Fact-to-occurrence anchoring predicate
$s_T, s_{TI}, s_{II}; \lambda_T, \lambda_{TI}, \lambda_{II}$	Fact-level component scores and weights	w_T, w_I, c_T, c_I	Entity-linking weights and similarities
s_F, s_Z	Fact- and evolution-level relevance scores	ϕ_T, ϕ_I, ψ_T	Text, image, and cross-modal encoders
$\text{Traverse}(v, \mathcal{E}', \mathcal{V}')$	Cross-layer traversal operator	R_z	Source rounds cited by evolution z
$\mathcal{R}, \text{Src}(z)$	Retained archive and recorded source set of z	K	Candidate limit in construction

B Additional Method Details

This section provides more detailed hypergraph schema, notation definition, construction and search algorithms, tool specifications, and a source-recoverability proof.

B.1 Detailed Hypergraph Memory Schema

MHG-Mem is an attributed hypergraph with three node types and three typed n-ary hyperedge families:

Nodes. FactNode is an atomic multimodal observation such as an event, state, visual attribute, or spatial relation. Each $f \in \mathcal{F}$ stores

$$f = (\ell_f, \mathcal{I}_f, t_f, s_f, \delta_f), \tag{12}$$

where ℓ_f is the fact description, $\mathcal{I}_f$ is its set of image identifiers, and (t_f, s_f, δ_f) gives the source round, session, and date.

EntityNode represents a distinct entity mention expressed as a text span, an image region, or both. Each $u \in \mathcal{U}$ stores

$$u = (n_u, c_u, \mu_u, i_u, b_u), \qquad b_u = (x_1, y_1, x_2, y_2) \in [0,1]^4, \tag{13}$$

where n_u is its name, c_u its entity class, μ_u its modality, i_u an optional image identifier, and b_u an optional normalized visual region. EntityNodes linked to the same underlying entity form a referent group r with mention set $\mathcal{U}_r$; every mention remains a distinct observation.

EvolutionNode records a change in an attribute or relation of a referent group, such as changes in appearance, price, or location. Each $z \in \mathcal{Z}$ stores

$$z = (r_z, a_z, \sigma_z, S_z, \kappa_z), \qquad S_z = ((v_j, f_j))_{j=1}^{m_z}, \qquad \mathcal{F}_z = \{f_j : 1 \le j \le m_z\} \subseteq \mathcal{F}_{r_z}, \tag{14}$$

where r_z is the subject referent, a_z is the evolving attribute or relation, σ_z summarizes the observed states, S_z is the ordered state–fact trace, and κ_z is the update type.

Hyperedges. Occurrence hyperedge binds all mentions and associated facts of a referent group:

$$\iota(e_r^{\text{occ}}) = \mathcal{U}_r \cup \mathcal{F}_r, \qquad \mathcal{U}_r = \iota(e_r^{\text{occ}}) \cap \mathcal{U}. \tag{15}$$

Evolutionary hyperedge binds an EvolutionNode to its subject and coactor mentions:

$$\iota(e_z^{\text{evo}}) = \{z\} \cup \mathcal{U}_z^{\text{subj}} \cup \mathcal{U}_z^{\text{co}}, \tag{16}$$

where $\mathcal{U}_z^{\text{subj}} \subseteq \mathcal{U}_{r_z}$ contains mentions of the changing entity and $\mathcal{U}_z^{\text{co}}$ contains mentions from other participating referents.

Evidential hyperedge connects the evolution to its supporting facts:

$$\iota(e_z^{\text{evd}}) = \{z\} \cup \mathcal{F}_z. \tag{17}$$

Finally, $\rho(v)$ gives the source-round reference for $v \in \mathcal{F} \cup \mathcal{U}$, while $\mathcal{I}_f$ and (i_u, b_u) retain image-level and region-level grounding. The complete memory is

$$\mathcal{M}_T = (\mathcal{H}, \rho, \mathcal{K}). \tag{18}$$

B.2 Detailed Memory Construction

Algorithm 1 combines the three construction steps. Original rounds and images are archived before extraction. FactNodes retain round-level image references $\mathcal{I}_f$, whereas visual EntityNodes retain observation-specific regions (i_u, b_u); both use ρ to locate their source rounds.

Algorithm 1 MHG-Mem construction

Require: History $\mathcal{D}_{\leq T}$, construction models, candidate limit K
Ensure: Memory $\mathcal{M}_T = (\mathcal{H}, \rho, \mathcal{K})$
Initialize the archive, $\mathcal{H}$, ρ, and $\mathcal{K}$
for $t = 1, \ldots, T$ **do**
  Archive d_t; index its images and captions
  Extract $\mathcal{F}_t$ using Equation equation 3 ▷ **Step 1**
  Store $\mathcal{F}_t$, source references, and text embeddings
  Extract entity candidates from d_t ▷ **Step 2**
  for each candidate u **do**
    Rank occurrence hyperedges by Equation equation 4
    Condition on d_t, u, and the top-K candidates
    Select `UPDATE`, `ADD`, or `SKIP`
    if the action is not `SKIP` **then**
      Store u and $\rho(u)$; execute the occurrence update
      Index its name and available visual region
    end if
  end for
  Anchor $\mathcal{F}_t$ using Equation equation 5
end for
for each referent group r with facts across sessions **do**
  Generate $\widehat{\mathcal{Z}}_r$ using Equation equation 6 ▷ **Step 3**
  Filter and order state references; apply Equation equation 19
  for each retained $z \in \widehat{\mathcal{Z}}_r$ **do**
    Set $r_z = r$; construct subject and coactor memberships
    Store z, e_z^{evo}, and e_z^{evd}
    Index $\phi_T(a_z \parallel \sigma_z)$ in $\mathcal{K}$
  end for
end for
return $\mathcal{M}_T$

Evolution validation and participants. The synthesizer reads ordered fact descriptions and their identifiers, and proposes histories for changing attributes or relations across sessions. Invalid fact references and repeated references are removed; the remaining state–fact pairs are ordered by t_f. For a candidate with subject $r_z = r$, retention requires

$$\begin{aligned} & f_j \in \mathcal{F}_r && (1 \leq j \leq m_z), \\ & t_{f_j} \leq t_{f_{j+1}} && (1 \leq j < m_z), \\ & \big|\{v_j : 1 \leq j \leq m_z\}\big| \geq 2. \end{aligned} \tag{19}$$

The values must be nonempty, and distinctness is checked after case and whitespace normalization. These checks validate the recorded sequence; identifying a meaningful change remains the synthesizer's task.

Let $R_z = \{\rho(f) : f \in \mathcal{F}_z\}$ be the cited round references. Subject mentions are selected from $\mathcal{U}_{r_z}$ with $\rho(u) \in R_z$; if none occurs in these rounds, one stored subject mention identifies the referent. For each other group $r' \neq r_z$ with $\mathcal{F}_{r'} \cap \mathcal{F}_z \neq \varnothing$, a mention satisfying $\rho(u) \in R_z$ is selected as a coactor when available. These role sets define e_z^{evo}; the distinct state references define e_z^{evd} through Equations equation 16 and equation 17. The ordered correspondence remains in S_z.

Incremental construction. At a new session boundary, only groups receiving newly anchored facts require resynthesis. Their evolution records are regenerated from the updated $\mathcal{F}_r$. Successful replacements update the derived nodes, hyperedges, and embeddings while retaining the underlying facts, mentions, and original sources.

Table 4: Four tool types for hypergraph-based agentic search. Query images, when supplied, are available to the visual lookups.

Type	Tool	Input	Returned information
Hypergraph	`resolve_facts`	Text query w	Ranked facts and source references.
	`resolve_entity`	Entity name or query image	Candidate referent groups and corresponding occurrence linking facts.
	`resolve_evolution`	Referent name, query image, or `evolution_id`	Evolutions and evidential linking facts.
Visual Inspection	`search_image`	Text query or query image	Matching images.
	`get_raw_verbatim`	Source-round identifier	Original dialogue and images.
Assembly	`submit_evidence`	Fact and image references; optional summary	Validation feedback; accepted evidence forms E_q.

B.3 Agentic Search and Tool Specification

Table 4 specifies the six tools available to the controller. Algorithm 2 uses them under a step budget B_{step} and a budget B_{img} on distinct images presented during search.

Multimodal embedding scoring. Using the text encoder ϕ_T from the main text, let ϕ_I encode images and ψ_T encode text in ϕ_I's feature space. An image identifier is resolved to its pixels before encoding. The components of Equation equation 9 are

$$\begin{aligned} s_T(w,f) &= \cos\big(\phi_T(w), \phi_T(\ell_f)\big), \\ s_{TI}(w,f) &= \max\big(\{0\} \cup \{\cos(\psi_T(w), \phi_I(I)) : I \in \mathcal{I}_f\}\big), \\ s_{II}(q,f) &= \max\big(\{0\} \cup \{\cos(\phi_I(J), \phi_I(I)) : J \in \mathcal{I}_q,\ I \in \mathcal{I}_f\}\big). \end{aligned} \tag{20}$$

Visual scores are floored at zero, and unavailable terms contribute zero without renormalizing $\lambda_T, \lambda_{TI}, \lambda_{II}$. The Visual Tool searches stored captions for a text lookup and image embeddings for a query-image lookup.

Traversal and evidence assembly. Entity and evolution lookups perform the membership expansions described in Section 3.4. Evolution relevance uses $a_z \parallel \sigma_z$, while readout dereferences the supporting facts in the order retained by S_z. Source inspection follows $\rho(f)$ to recover original observations. Facts returned by retrieval or associated with accessed rounds are registered for citation. At submission, fact identifiers must be registered and available, and attached image identifiers must belong to the cited fact's source round. Invalid references are removed. The retained references select textual and visual evidence for E_q, together with the optional controller summary. These checks validate identifier associations, not the semantic content of that summary.

B.4 Provenance Preservation

Let $\mathcal{R}$ be the archive mapping source-round identifiers to retained observations, so $\rho(f) \in \mathrm{dom}(\mathcal{R})$ means that the source of f is available. For $S_z = ((v_j, f_j))_{j=1}^{m_z}$, define the recorded source set

$$\mathrm{Src}(z) = \{\rho(f_j) : 1 \le j \le m_z\}. \tag{21}$$

The following property concerns access to recorded evidence, independently of extraction accuracy or the truth of generated summaries.

Proposition 1 (Source recoverability)**.** *Suppose every retained state cites $f_j \in \mathcal{F}_{r_z}$ with $\rho(f_j) \in \mathrm{dom}(\mathcal{R})$, and e_z^{evd} stores exactly the distinct facts cited in S_z as in Equation equation 17. Then each state has a trace to its original observation, and expanding e_z^{evd} followed*

by ρ recovers exactly $\mathrm{Src}(z)$*. This property is preserved by fact and mention additions and validated evolution insertions or replacements with valid fact identifiers, source bindings, and archive records.*

Proof. Because $\mathcal{F}$ and $\mathcal{Z}$ are disjoint, $z \notin \mathcal{F}$. Since every $f_j \in \mathcal{F}_{r_z} \subseteq \mathcal{F}$, Equation equation 11 gives

$$\begin{aligned} \mathrm{Traverse}(z, \{e_z^{\mathrm{evd}}\}, \mathcal{F}) &= \iota(e_z^{\mathrm{evd}}) \cap \mathcal{F} \\ &= (\{z\} \cup \mathcal{F}_z) \cap \mathcal{F} = \mathcal{F}_z. \end{aligned} \tag{22}$$

For every $1 \le j \le m_z$, the membership $f_j \in \mathcal{F}_z$ and the source-availability assumption therefore define the trace

$$z \xrightarrow{e_z^{\mathrm{evd}}} f_j \xrightarrow{\rho} \rho(f_j) \xrightarrow{\mathcal{R}} \mathcal{R}(\rho(f_j)). \tag{23}$$

Applying ρ to all recovered fact members yields

$$\begin{aligned} &\{\rho(f) : f \in \mathrm{Traverse}(z, \{e_z^{\mathrm{evd}}\}, \mathcal{F})\} \\ &\quad = \{\rho(f) : f \in \mathcal{F}_z\} \\ &\quad = \{\rho(f_j) : 1 \le j \le m_z\} = \mathrm{Src}(z). \end{aligned} \tag{24}$$

Thus expansion neither adds nor omits a recorded source. The set equality deduplicates shared sources; the index-specific association $j \mapsto f_j \mapsto \rho(f_j)$ remains available through S_z.

For preservation, denote the evolution sets before and after an update by $\mathcal{Z}^-$ and $\mathcal{Z}^+$. If $\mathcal{Z}_{\mathrm{del}}$ contains removed evolutions and $\mathcal{Z}_{\mathrm{new}}$ contains validated new evolutions, then

$$\mathcal{Z}^+ = (\mathcal{Z}^- \setminus \mathcal{Z}_{\mathrm{del}}) \dot{\cup} \mathcal{Z}_{\mathrm{new}}. \tag{25}$$

For every surviving z, unchanged state references and source bindings give

$$\mathrm{Src}^+(z) = \mathrm{Src}^-(z) \subseteq \mathrm{dom}(\mathcal{R}^-) \subseteq \mathrm{dom}(\mathcal{R}^+), \tag{26}$$

where superscripts identify the corresponding memory snapshots. Its evidential membership is unchanged, so Equations equation 22–equation 24 continue to hold. For every new evolution, validation supplies the same assumptions, and the preceding argument applies directly. Fact and mention additions are the case $\mathcal{Z}_{\mathrm{del}} = \mathcal{Z}_{\mathrm{new}} = \varnothing$. The empty memory satisfies the property vacuously; induction over these operations establishes it for every successfully constructed snapshot. □

C Additional Experimental Details

C.1 Benchmarks and Metrics

MemEye. MemEye (Guo et al., 2026) evaluates visual memory along two axes: visual evidence granularity, from scene-level to pixel-level details, and memory reasoning depth, from atomic retrieval to evolutionary synthesis. It contains eight everyday scenarios and 371 questions, each with paired multiple-choice and open-ended versions. These tasks require agents to recover visually grounded facts, distinguish similar instances, and reason over observations accumulated across sessions.

Mem-Gallery. Mem-Gallery (Bei et al., 2026) evaluates long-term multimodal conversational memory using persona-driven dialogues across 20 scenarios. The evaluated release contains 240 sessions, 3,962 dialogue rounds, 1,490 images, and 1,711 open-ended questions. It probes memory extraction, reasoning, and knowledge management over extended text–image histories.

Algorithm 2 Hypergraph-based agentic search

Require: Query q, memory $\mathcal{M}_T$, controller π_θ, step budget B_{step}, image budget B_{img}
Ensure: Retrieved evidence E_q

```
Initialize h_1 = () and an empty registry of retrieved facts
for k = 1, ..., B_step do
    Reason over q, h_k; select a_k ~ π_θ(· | q, h_k)
    if a_k selects hypergraph retrieval then
        o_k ← layer-specific retrieval with applicable traversal (Eq. equation 11)
    else if a_k selects visual retrieval then
        o_k ← matching images and captions
    else if a_k selects source inspection then
        o_k ← original dialogue and images via ρ
    else                                                    ▷ Evidence submission
        Validate submitted references and source associations
        if submission is accepted then
            return E_q assembled from selected evidence and notes
        end if
        o_k ← feedback requesting additional evidence
    end if
    Register retrieved facts and facts associated with accessed sources
    Attach requested images to o_k within B_img
    h_{k+1} ← h_k ‖ (a_k, o_k)
end for
return Finalize E_q through forced evidence submission
```

WorldMemArena-S. WorldMemArena (Liu et al., 2026a) evaluates memory in GUI and embodied-agent trajectories, as well as lifelong personal and project interactions. We use its released 150-sample subset, denoted WorldMemArena-S, containing 7,906 questions. Its 11 question types cover textual memory, visual memory, and reasoning, including factual recall, dynamic updates, conflict resolution, and cross-modal reasoning. Evaluation follows the annotated QA checkpoints: MHG-Mem constructs memory from only the sessions covered by each checkpoint. Prefix checkpoints reuse the corresponding historical state, while non-prefix session windows are constructed independently to prevent access to out-of-scope observations.

Evaluation Metrics. Let $\mathcal{Q}_b$ denote the open-ended questions of benchmark b and $\mathcal{Q}^{\text{MCQ}}_{\text{MemEye}}$ its multiple-choice questions. Answers are lowercased, punctuation-normalized, and whitespace-normalized before scoring; all reported values are percentages.

For MemEye, options are rotated across R_q equivalent instances. Exact Match (EM) is the mean accuracy over rotations and questions:

$$\text{EM} = \frac{1}{|\mathcal{Q}^{\text{MCQ}}_{\text{MemEye}}|} \sum_q \frac{1}{R_q} \sum_{r=1}^{R_q} \mathbf{1}[\hat{c}_{q,r} = c^\star_{q,r}]. \tag{27}$$

For open-ended answers, let $\widetilde{T}(\cdot)$ be the normalized token multiset after Porter stemming and $C_q = \widetilde{T}(\hat{y}_q) \cap \widetilde{T}(y^\star_q)$. Token F1 is

$$F_{1,q} = \frac{2P_qR_q}{P_q + R_q}, \quad P_q = \frac{|C_q|}{|\widetilde{T}(\hat{y}_q)|}, \quad R_q = \frac{|C_q|}{|\widetilde{T}(y^\star_q)|}, \qquad \text{F1}_b = \frac{1}{|\mathcal{Q}_b|} \sum_{q \in \mathcal{Q}_b} F_{1,q} \tag{28}$$

with $F_{1,q} = 0$ when the overlap is empty. BLEU-1 uses the same normalization without stemming and combines clipped unigram precision with a brevity penalty:

$$\text{BLEU1}_q = \text{BP}_q \frac{\sum_w \min\{c_{\hat{y}_q}(w), c_{y^\star_q}(w)\}}{\sum_w c_{\hat{y}_q}(w)}, \qquad \text{B-1}_b = \frac{1}{|\mathcal{Q}_b|} \sum_{q \in \mathcal{Q}_b} \text{BLEU1}_q \tag{29}$$

where $\mathrm{BP}_q = 1$ if $|\hat{y}_q| > |y_q^\star|$ and otherwise $\mathrm{BP}_q = \exp(1 - |y_q^\star|/|\hat{y}_q|)$. We use NLTK method-1 smoothing for sentence-level BLEU-1.

MemEye and Mem-Gallery use their native reference-based judge, whose parsed score is $s_q \in \{0, 0.5, 1\}$; $\mathrm{Judge}_b = |\mathcal{Q}_b|^{-1} \sum_{q \in \mathcal{Q}_b} s_q$. WorldMemArena-S labels each answer as correct, hallucination, or omission and reports QA-C as its Judge:

$$\text{QA-C} = \mathrm{Judge}_{\text{WMA-S}} = \frac{|\{q \in \mathcal{Q}_{\text{WMA-S}} : \ell_q = \text{correct}\}|}{N_{\text{valid}}} \tag{30}$$

For WorldMemArena memory diagnostics, let $\mathcal{G}_s$ be the gold memory points, $\mathcal{D}_s$ the stored add/update items, and $\mathcal{C}_s$ the gold points judged covered by $\mathcal{D}_s$. With importance weight w_g, Memory Recall is

$$\texttt{Memory_Recall}_s = \frac{\sum_{g \in \mathcal{C}_s} w_g}{\sum_{g \in \mathcal{G}_s} w_g}, \qquad \texttt{Memory_Recall} = \mathrm{mean}_{s:\,|\mathcal{G}_s|>0}\, \texttt{Memory_Recall}_s. \tag{31}$$

If n_s^{C} stored items are judged correct, Memory Correctness is

$$\texttt{Memory_Correctness} = \mathrm{mean}_{s:\,|\mathcal{D}_s|>0}\, \frac{n_s^{\mathrm{C}}}{|\mathcal{D}_s|}, \tag{32}$$

implemented by the released evaluator as $1 - n_s^{\text{incorrect}}/n_s^{\text{checked}}$. For each update, the post-session memory is classified as updated, both, or outdated, yielding

$$\texttt{Update_Handling} = \frac{N_{\text{updated}} + 0.5 N_{\text{both}} + 0 N_{\text{outdated}}}{N_{\text{total}}}. \tag{33}$$

Finally, if $\mathcal{V}_q$ is the gold evidence set and covered_q is the number of evidence points found by the LLM judge in the returned retrieval list, retrieval hit rate is

$$\texttt{retrieval_hit_rate} = \frac{1}{|\mathcal{Q}_{\text{WMA-S}}|} \sum_{q \in \mathcal{Q}_{\text{WMA-S}}} \frac{\text{covered}_q}{|\mathcal{V}_q|}, \tag{34}$$

where questions without gold evidence contribute zero. WMA metrics are first computed per trajectory and then averaged across evaluated instances.

Table 5 consolidates the fine-grained task semantics used by Mem-Gallery and WorldMemArena-S. Shared capabilities are listed once, while benchmark-specific labels and definitions are retained.

C.2 Implementation Details

We use Qwen3-VL-8B, GPT-4.1-mini, and Qwen3.8-27B memory backbones. The selected backbone performs memory construction, agentic retrieval,and answer generation in MHG-Mem. We instantiate the text encoder ϕ_T with `all-MiniLM-L6-v2`, and the image and cross-modal encoders (ϕ_I, ψ_T) with `google/siglip2-base-patch16-384`. Qwen3-VL-8B and Qwen3.8-27B are served using vLLM on four NVIDIA A800 GPUs, while GPT-4.1-mini is accessed through the official API. All models apply 65,536 max token limit and 0 temperature. Each result is averaged over three runs.

In memory construction stage, the context $\mathcal{C}_t$ contains up to three preceding rounds from the same session, and at most eight images from $\mathcal{I}_t$ are passed to the construction models. Occurrence linking uses the top $K = 3$ candidate hyperedges in Eq. equation 4, with $w_T = w_I = 0.5$. Evolution synthesis retains only records z grounded in at least two facts in $\mathcal{F}_z$ and observations from at least two sessions. For WorldMemArena-S, $\widehat{\mathcal{Z}}_r$ is refreshed after every visible session checkpoint; for the static MemEye and Mem-Gallery corpora, it is synthesized after the available history $\mathcal{D}_{\le T}$ has been constructed.

In memory retrieval stage, the controller π_θ follows the search loop in Algorithm 2 with $B_{\text{step}} = 16$ and $B_{\text{img}} = 10$. The entity-layer search returns at most 3 referent groups; the

Table 5: Fine-grained task definitions used by Mem-Gallery and WorldMemArena-S. A dash indicates that the benchmark has no directly corresponding task label.

Capability	Mem-Gallery	WorldMemArena-S
Fact recall	**FR** (Factual Retrieval): retrieve an explicit fact from the conversation.	**FR** (Fact Recall): retrieve one concrete fact stated earlier in the trajectory.
Multi-entity reasoning	**MR** (Multi-entity Reasoning): combine retrieved information about multiple entities to infer an answer.	–
Temporal reasoning	**TR** (Temporal Reasoning): resolve a time-dependent question.	**TR** (Temporal Reasoning): reason about timing, ordering, or duration.
Visual or cross-modal reasoning	**VR** (Visual-centric Reasoning): use visual information in addition to the dialogue.	**CMR** (Cross-modal Reasoning): combine textual and visual memory.
Test-time learning	**TTL** (Test-time Learning): learn visual knowledge from historical images and apply it during QA.	**TTL** (Test-Time Learning): apply a rule or skill taught earlier in the trajectory.
Visual search	**VS** (Visual-centric Search): find the image(s) matching a query and return their image IDs.	**VS** (Visual Search): locate an object or attribute across visual memory.
Conflict handling	**CD** (Conflict Detection): detect a contradiction between the history and the question.	**MC** (Memory Conflict): resolve disagreement between two memory items using context.
Knowledge reasoning or resolution	**KR** (Knowledge Resolution): resolve updates or conflicts by prioritizing the most recent information.	**KR** (Knowledge Reasoning): combine stored facts with general world knowledge.
Answer boundary	**AR** (Answer Refusal): answer "Not mentioned" when the information is absent.	–
Dynamic update	–	**DU** (Dynamic Update): report the latest value after a queried fact is overwritten.
Visual fact recall	–	**VFR** (Visual Fact Recall): recover a fact anchored to a specific image.
Visual update	–	**VU** (Visual Update): report the most recent state after a visual attribute changes.

visual search returns at most 5 images. Evolution-layer search returns at most 3 relevant evolution aspects. The fact-layer search returns at most 30 candidates facts. For MemEye and Mem-Gallery, source inspection exposes the rounds referenced by $\rho(f)$; WorldMemArena-S exposes the cited FactNodes $f \in \mathcal{F}$. We follow the benchmark-specific QA and judging procedures defined above.

C.3 Baselines

A-Mem. A-Mem (Xu et al., 2025) organizes textual memories as interconnected notes inspired by the Zettelkasten method. It enriches each note with contextual descriptions and indexing attributes, establishes links to related notes, and updates existing memories as new observations arrive.

MemoryOS. MemoryOS (Kang et al., 2025) maintains short-term, mid-term, and long-term memory. Storage, updating, retrieval, and generation modules coordinate the transfer and consolidation of conversational information across these tiers to support persistent user knowledge and preferences.

SimpleMem. SimpleMem (Liu et al., 2026c) compresses interactions into compact, self-contained textual memory units. Its pipeline combines semantic structured compression, online semantic synthesis, and intent-aware retrieval planning to reduce redundancy and assemble query-relevant context. The unimodal category refers to textual memory representations.

HyperMem. HyperMem (Yue et al., 2026) organizes long-term conversations into topic, episode, and fact levels. Hyperedges group associated episodes and facts, while hybrid lexical–semantic indexing and coarse-to-fine retrieval recover related information across these levels.

M2A. M2A (Feng et al., 2026) combines an immutable raw-message store with a semantic-memory store of higher-level observations. A conversational agent and a memory manager coordinate memory updates and iterative retrieval, retaining links between semantic memories and their original multimodal evidence.

MMA. MMA (Lu et al., 2026) is a confidence-aware multimodal memory agent. It combines semantic relevance with memory reliability, estimated from source credibility, temporal decay, and conflict-aware consensus, to reduce reliance on stale or conflicting observations.

SRAG. SRAG denotes the multimodal Semantic RAG baseline in MemEye (Guo et al., 2026). It ranks historical records by combining text and image embedding similarities and supplies retrieved text and associated original images to the answer model. It provides a direct retrieval comparison to methods that construct additional memory abstractions or relational structures.

UniversalRAG. UniversalRAG (Yeo et al., 2026) routes queries to corpora organized by modality and granularity. Modality-aware routing selects the appropriate evidence source, while retrieval at different granularities supplies context for answer generation.

D Additional Experiment Results and Analysis

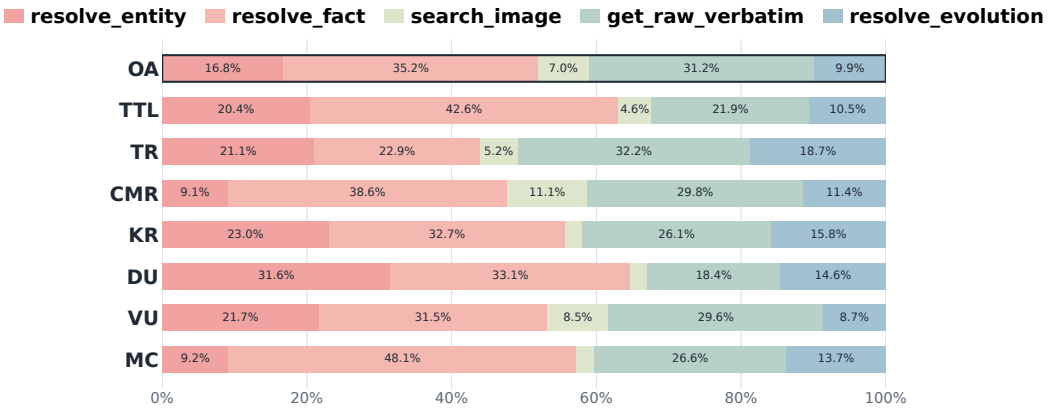

Figure 5: Average tool utilization on WorldMemArena-S with Qwen3.8-27B, reported overall and by 7 task types: OA (Overall Average), TTL (Test-time Learning), TR (Temporal Reasoning), CMR (Cross-modal Reasoning), KR (Knowledge Reasoning), DU (Dynamic Updating), VU (Visual Update), and MC (Memory Conflict).

D.1 Retrieval Tool Analysis

Figure 5 reports average tool usage on WorldMemArena-S with Qwen3.8-27B. Hypergraph tools (`resolve_entity`, `resolve_fact`, and `resolve_evolution`) and the inspection tool (`get_raw_verbatim`) dominate agentic search. In particular, `resolve_fact` and `get_raw_verbatim` are invoked most frequently, suggesting that many questions can be

traced through atomic facts and verified against original dialogue or visual evidence. Entity- and evolution-layer retrieval also accounts for substantial usage in TTL, TR, KR, DU, and VU tasks. This pattern suggests that, for temporal reasoning and multimodal updates, the agent links observations of the same entity and retrieves relevant state histories to reason about attribute updates, conflicts, and replacements.

D.2 Cost Analysis

Figure 6 compares token consumption and QA performance across methods. Following prior work, we measure the average total number of LLM input and output tokens per sample across memory construction, retrieval, and QA. For multimodal methods, this count also includes image input tokens determined by the corresponding LLM's visual encoding. We additionally report MHG-Mem (w/o A.S.), an ablation that replaces agentic search with top-k hypergraph retrieval.

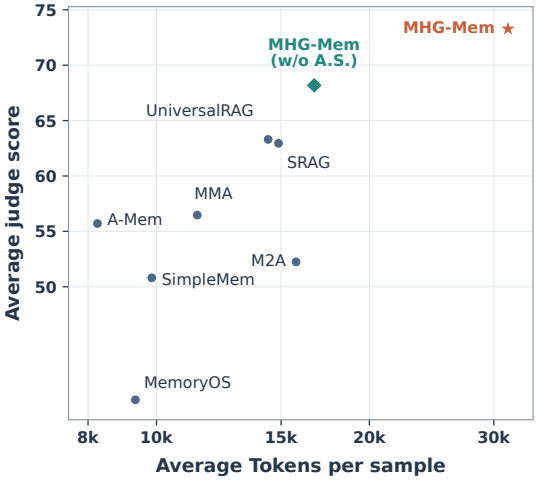

Figure 6: Trade-off between token consumption and QA performance among different baselines. w/o A.S. represents replace MHG-Mem's agentic search with top-k retrieval.

Among the baselines, unimodal memory methods that abstract images into textual captions consume fewer tokens but achieve lower QA performance, as fine-grained visual details are omitted. Among multimodal methods, RAG-based approaches (SRAG, MMA, and UniversalRAG) require fewer tokens than memory-based approaches (M2A and MHG-Mem) because they avoid explicit memory construction and iterative search. Although MHG-Mem improves benchmark-averaged QA performance by 8.20–9.98 points over the strongest baseline for each backbone, it consumes approximately 2x tokens, reflecting both the processing of multimodal information during memory construction and iterative tool calls during retrieval. The additional cost primarily arises from agentic search: MHG-Mem (w/o A.S.) requires only approximately 1–2k more tokens than multimodal baselines while still delivering performance gains of 5%–15%, offering a favorable balance between cost and performance. These results suggest using similarity-based hypergraph retrieval in resource-constrained settings and agentic hypergraph search when higher QA performance is the priority.

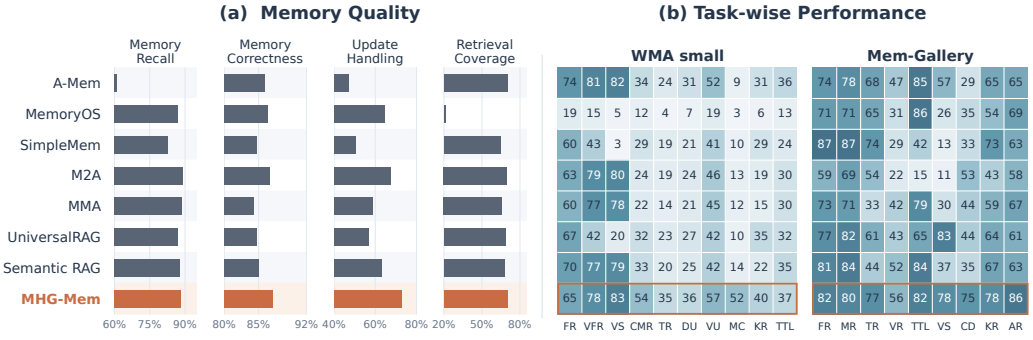

Figure 7: (a) Memory quality and (b) fine-grained tasks performance on WorldMemArena-S (WMA-S) and Mem-Gallery with Qwen3-VL-8B.

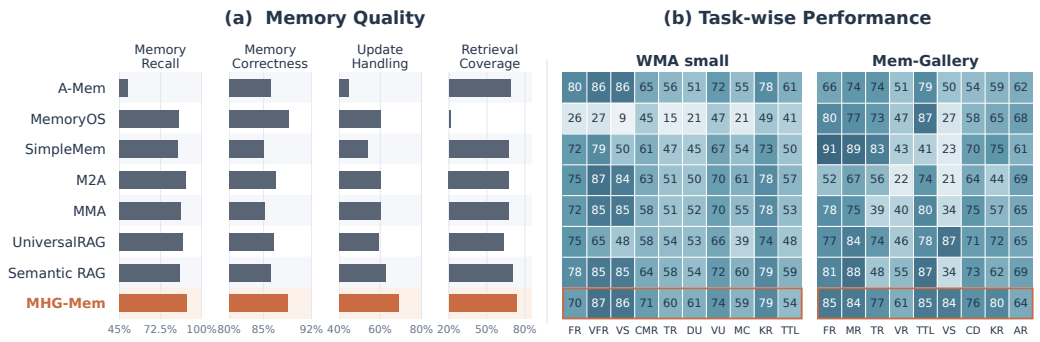

Figure 8: (a) Memory quality and (b) fine-grained tasks performance on WorldMemArena-S (WMA-S) and Mem-Gallery with GPT-4.1-mini.

D.3 Memory Quality and Fine-grained Analysis on More Backbones

Figure 7 and 8 extend the memory quality and task-level analysis to Qwen3-VL-8B and GPT-4.1-mini, respectively. Across both backbones, MHG-Mem achieves the highest memory correctness and update handling, with retrieval coverage among the best. These results show that effective state maintenance coexists with accurate memory retention and strong evidence coverage, consistent with our objective of organizing evolving observations while keeping their supporting evidence accessible.

The most consistent task-level advantages concern visual evidence use and state evolution. On WorldMemArena-S, MHG-Mem outperforms all compared baselines in cross-modal reasoning, temporal reasoning, dynamic updates, and visual updates with both backbones. In particular, cross-modal reasoning improves over the strongest baseline by 20% and 6% with Qwen3-VL-8B and GPT-4.1-mini, respectively. On Mem-Gallery, visual-centric reasoning improves by 4% and 6%, while knowledge resolution improves by 5% under both backbones; conflict detection also benefits in both settings. These recurring gains align with the two requirements motivating MHG-Mem: accessing relevant visual observations and integrating changes across interaction histories. They are consistent with the intended role of entity histories and state-specific evidence links in supporting comparison and verification across observations.

The improvements nevertheless depend on the task and backbone. For example, Mem-Gallery temporal reasoning improves by 3% with Qwen3-VL-8B but remains 6% below the strongest baseline with GPT-4.1-mini. MHG-Mem also trails the strongest factual-recall baseline on both benchmarks under both backbones. Thus, the memory-quality improvements do not translate uniformly into every downstream capability: visual evidence use and state updates show reproducible benefits, while factual recall remain areas for improvement.

D.4 Hyperparameter Analysis

Analysis of the occurrence-linking candidate count K. The upper half of Table 6 reports QA performance with different top-K candidate counts for occurrence linking during memory construction. Across the tested values of K, scores vary by only 1.48, 2.26, and 1.94 points on MemEye, Mem-Gallery, and WorldMemArena-S, respectively. The average score is 73.28 at $K = 3$, compared with 73.33 at the best tested value, $K = 5$, indicating that performance is relatively insensitive to K. Since larger candidate sets increase context length with limited performance gains, we retrieve at most $K = 3$ candidates for occurrence linking.

Analysis of agentic budgets B_{step} and B_{img}. The lower half of Table 6 examines the effects of retrieval budgets. Increasing B_{step} or B_{img} allows longer retrieval trajectories or more image evidence, respectively, but also increases input token consumption. Raising the budgets from $(4, 4)$ to $(16, 8)$ improves the average QA score from 67.29 to 72.48, a

Table 6: Hyperparameter sensitivity at the construction and retrieval stages with Qwen3.8-27B. The construction-stage sweep varies the number of top-K occurrence-linking candidates. The retrieval-stage sweep evaluates the agentic search budget $(B_{\text{step}}, B_{\text{img}})$ and reports the force-submit rate. LLM-Judge scores are reported. Avg. means the average judge score across three benchmark. Bold values are the best within each stage.

Setting	MemEye	Mem-Gallery	WMA-S	Avg.	Force-submit (%)
Construction stage					
$K = 1$	64.70	85.84	64.41	71.65	—
$K = 2$	65.54	86.99	65.67	72.73	—
$K = 3$	65.93	87.85	66.05	73.28	—
$K = 4$	65.22	87.46	**66.35**	73.01	—
$K = 5$	**66.18**	**88.10**	65.71	**73.33**	—
Retrieval stage					
$(B_{\text{step}}, B_{\text{img}}) = (4, 4)$	59.44	81.79	60.63	67.29	37.3
$(B_{\text{step}}, B_{\text{img}}) = (10, 4)$	62.02	84.76	62.37	69.72	26.8
$(B_{\text{step}}, B_{\text{img}}) = (10, 8)$	63.91	86.30	63.10	71.10	24.6
$(B_{\text{step}}, B_{\text{img}}) = (16, 8)$	65.42	87.04	64.99	72.48	17.9
$(B_{\text{step}}, B_{\text{img}}) = (16, 10)$	65.73	**87.85**	66.05	73.21	17.5
$(B_{\text{step}}, B_{\text{img}}) = (24, 10)$	65.47	87.63	66.51	73.20	**13.4**
$(B_{\text{step}}, B_{\text{img}}) = (24, 16)$	**66.09**	87.46	**66.80**	**73.45**	13.5

gain of 5.19 percentage points, while reducing the forced-submission rate from 37.3% to 17.9%. These results suggest that restrictive budgets can prevent the agent from gathering sufficient evidence through hypergraph search. For example, questions requiring fine-grained visual distinctions among multiple entities may require inspecting more image crops than the budgets permit, leading to premature submission and lower performance. However, increasing $(B_{\text{step}}, B_{\text{img}})$ further to $(24, 16)$ yields only a 0.24-point improvement over $(16, 10)$ while consuming more tokens. To balance the cost and performance trade-off, we therefore set $(B_{\text{step}}, B_{\text{img}}) \leq (16, 10)$ for experiment.

D.5 More Case Studies

Figures 9 and 10 illustrate how MHG-Mem reconstructs entity histories through adaptive retrieval and source verification. In Figure 9, the agent retrieves the faucet's evolution record and traces its supporting facts to original frames, recovering the complete *off–running–off* sequence; the baselines omit states or fail to answer. In Figure 10, retrieval begins with entity resolution, followed by evolution retrieval and source inspection. This recovers the oak-log count sequence $1 \rightarrow 2 \rightarrow 5$ and verifies the latest count from the inventory image, whereas UniversalRAG and M2A recover only $1 \rightarrow 2$. Both cases highlight the value of linking observations under a shared entity identity while preserving state-specific evidence for cross-session synthesis and visual verification.

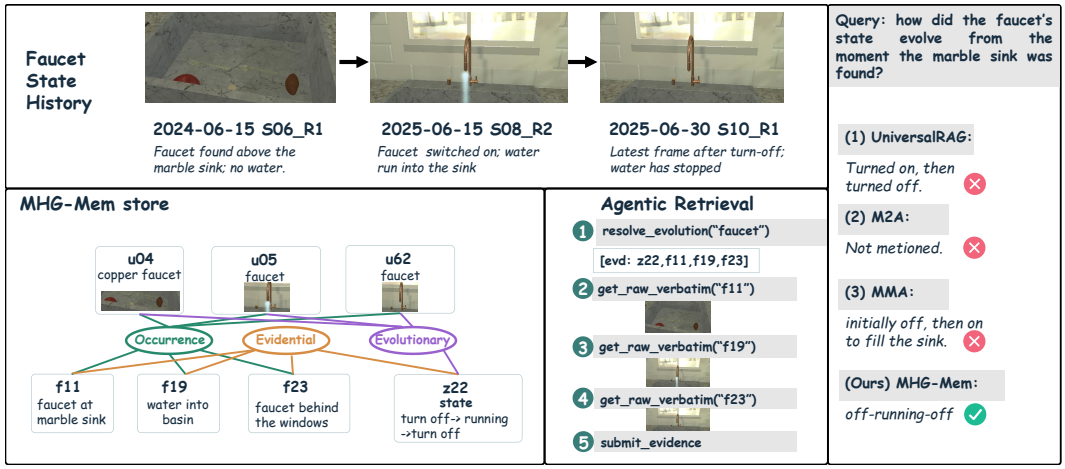

Figure 9: Case study in an embodied agent scenario. MHG-Mem records the faucet's state throughout the multimodal agent's action history.

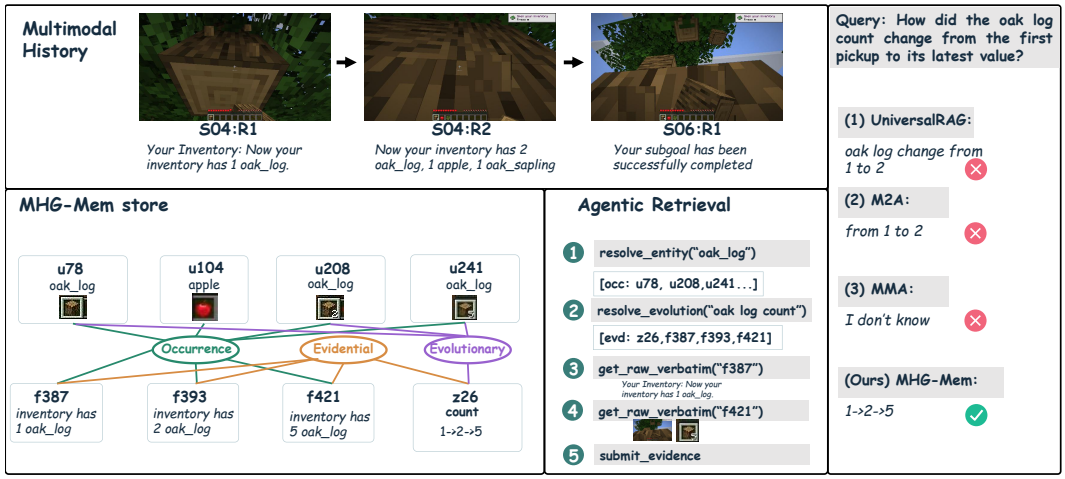

Figure 10: Case study in a GUI agent scenario. MHG-Mem records the count state of oak log in Minecraft throughout the multimodal agent's action history.

E Prompt Templates

The five prompt templates below cover memory construction (Appendix B.2) and agentic search (Table 4). We preserve prompt wording and placeholders, omit Python assignment wrappers, and adjust line wrapping for readability. For additional LLM-as-a-judge and QA prompts, we use the original benchmark-provided prompts to ensure fair comparisons.

E.1 Atomic fact construction

Converts the current round into standalone facts while retaining temporal and visual details.

01 Atomic fact construction

You convert ONE conversation round into a set of ATOMIC MEMORIES — short, standalone facts that a separate answerer will later retrieve to answer questions. You are given the round's dialogue text, its image caption(s), and its REAL image(s). You may also see up to three EARLIER rounds of the same conversation as read-only context (their text and image captions only — no earlier pictures), and a list of EXISTING MEMORIES already stored from past rounds. Write memories ONLY for the CURRENT round; use the earlier rounds and the current image(s) just to understand the round and to resolve references.

Read BOTH the caption AND the real pixels of the current round. The caption often spells a detail out in words (card names, on-screen text, dashboard labels, message transcripts); the pixels carry details the caption omits (colors, counts, spatial layout, what is actually shown). Capture from both.

ADD-ONLY against EXISTING MEMORIES

If an `## EXISTING MEMORIES` list is given, treat it as memories that are ALREADY STORED. Your job is ADDITIVE: output the facts from the CURRENT round that are NOT already represented there. Do not restate, paraphrase, or lightly reword an existing memory. Never create a memory from the existing list alone — every memory you output must be supported by the CURRENT round.

- Default to WRITING a fact. Suppress a candidate ONLY when it is the SAME fact about the SAME moment as an existing memory. When in doubt, WRITE it — a missed fact can never be recovered, a near-duplicate is cheap.

- The mere presence of the same entity, brand, game, route, dashboard, or template in an existing memory does NOT make the current round a duplicate. A new event, a new state, a changed value, a different date/session, a new opinion or conclusion the user draws, or any new detail is ALWAYS a new memory — write it.

Write each memory to this STANDARD:

- ATOMIC & self-contained: one fact per memory — a single event, state, attribute, or relationship — understandable on its own. Begin from the concrete named subject.

- ABSOLUTE time: anchor every time reference to the Observation Date given. Convert "today / tomorrow / yesterday / last week / this Monday" to an explicit calendar date (YYYY-MM-DD). Never store a bare relative time or weekday.

- DESCRIBE the state, do NOT compute the answer. Record what is shown or said — list the cards in a hand, the panels on a dashboard and their values, the people in a chat and who said what. Do NOT add totals, counts across rounds, winners, comparisons, or any conclusion the data merely implies. Faithful state, never a frozen verdict.

- CAPTURE what a later question could ask: numbers, colors, spatial relations (left / right / above / below / lower-left / behind / beside / under), in-image text, per-item attributes, who did what, and any change from the earlier context.

- Only return an empty list when the CURRENT round genuinely contains NOTHING beyond what is already stored (e.g. a bare "ok", "thanks", "correct" with no new content of its own).

Return ONLY JSON:

{"atomic_facts":["<one standalone memory>","<another standalone memory>"]}

If the round carries nothing new worth remembering, return {"atomic_facts":[]}.

E.2 Multimodal entity extraction

Extracts nameable entities and grounds visual or multimodal mentions in image regions.

02 Multimodal entity extraction

You extract MULTIMODAL ENTITIES from ONE conversation round (its dialogue text AND its image, if any).

An ENTITY is a persistent, nameable THING that could be asked about later — for example: a brand, logo, product. This list is illustrative, not exhaustive — extract any concrete, nameable thing a later question could point at

Give each entity ONE modality, decided by WHERE you learned about it:

- `visual`: something you can SEE in the image. This is the DEFAULT for anything in the picture — objects, people, places, logos, products, signs, on-screen panels. Give it a short specific name (e.g. "red delivery truck", "Coca-Cola logo"). bbox required.

- `multimodal`: visible in the image AND also named in the dialogue text. Name it from the dialogue; the box is on the region that shows it. bbox required.

- `text`: named in the dialogue text (what the USER or ASSISTANT actually typed) but NOT visible in the image. bbox is null. This is RARE — use it ONLY for things mentioned in the conversation that you cannot see.

For every `visual` and `multimodal` entity, bbox is absolute pixels `[x0,y0,x1,y1]` in the image's pixel frame, tight around the thing.

Return ONLY JSON:

{"entities":[{"name":"...","type":"<brand|person|character|object|product|place|animal|organization|other>","modality":"visual|multimodal|text","bbox":[x0,y0,x1,y1] or null}]}

Rules:

- Only extract things written in the dialogue or clearly visible in the image — never invent.

- If nothing qualifies, return {"entities":[]}.

E.3 Occurrence linking

Associates a candidate mention with a stored referent, creates a new referent, or rejects an invalid extraction.

03 Occurrence linking

You decide how to store ONE candidate entity into a memory of already-known entities.

You are given:

- ROUND CONTEXT: the conversation round the candidate was found in (its text, and the full round image if any) — use it to understand what the candidate actually is.

- CANDIDATE: its name, type, modality, and, if visual, its cropped region (shown below).

- MATCHES: most similar to the candidate (by text and visual similarity), each with its best-matching stored region (shown below).

Choose EXACTLY ONE action:

- `update`: the candidate is the SAME real-world thing as one MATCH— link to that entity id. Also use `update` for a repeat sighting of a thing already stored, to record this new occurrence.

- `add`: the candidate is a REAL entity but none of the MATCHES is the same real-world thing.

- `skip`: ONLY when the candidate is NOT a real, storable entity (an extraction mistake or abstract concept).

Return ONLY JSON:

`{{"action":"update|add|skip","target_entity_id":"<id, only if update>","add_alias":"<optional new alias>","reason":"<one line>"}}`

E.4 State evolution synthesis

Identifies changing aspects and binds each reported state to its supporting fact identifier.

04 State evolution synthesis

You are given ONE referent (a thing that recurs in a long conversation) and the time-ordered list of ATOMIC FACTS recorded about it. Your job is to report which ASPECTS of this referent actually CHANGED over time.

An ASPECT is a single named property that can hold different values at different times — for example: hand_size, location, owner, color, price, status, score, version, participants, weight.

What counts as a change

A change means: the SAME aspect of the SAME referent holds DIFFERENT values at different times. For example:

- "is in the kitchen" then "is in the bedroom"-> aspect `location` CHANGED. Report it.

- "costs $12" then "costs $9"-> aspect `price` CHANGED. Report it.

What does NOT count as a change (this is the common case — read carefully)

- A value stated repeatedly, or restated in different words, is NOT a change.

- New, additional details about DIFFERENT aspects are NOT a change unless one aspect's own value is replaced by another value.

How to write what you find

For each aspect that genuinely changed:

- `aspect`: a short lowercase snake_case key naming the property (`hand_size`, not "the size of the hand").

- `states`: two or more entries in TIME ORDER. Each entry MUST carry:

- `value`: the value at that time, as short as possible (a number, a color, a place, a name).

- `fact_id`: the id of the fact you read that value from. It MUST be one of the fact ids listed below, copied exactly. Never invent an id.

- `summary`: one sentence DESCRIBING the observed states in order, with their dates.

- `update_type`: one of

- `replacement` — the new value replaces the old one (a measurement, a location, a price)

- `correction` — the later fact says the earlier one was wrong

- `addition` — the value grew by accumulating items rather than being replaced

Return ONLY JSON:

{"aspects":[{"aspect":"<snake_case>","states":[{"value":"<v1>","fact_id":"<id>"},{"value":"<v2>","fact_id":"<id>"}],"summary":"<one sentence describing the states in order>","update_type":"replacement|correction|addition"}]}

If no aspect of this referent changed, return exactly {"aspects":[]}.

E.5 Hypergraph-based agentic search

Guides sequential tool use to assemble evidence for the separate answer model.

The placeholders `[MEMORY_SCHEMA]`, `[TOOL_SCHEMA]`, and `{max_steps}` denote the memory schema, tool definitions, and search budget, respectively. The five gathering tools collect evidence; `submit_evidence` terminates retrieval.

05 Hypergraph-based agentic search

You are a hypergraph memory retriever. Given a QUESTION, you SEARCH a long-term multimodal memory of past conversation rounds and assemble an EVIDENCE CHAIN for a separate QA model. You do NOT answer the question yourself.

What is in memory: [MEMORY_SCHEMA]

Tool Definition: [TOOL_SCHEMA]

Format — EXACTLY ONE action per reply, nothing after it

Each reply MUST be exactly:

```
Action: <one tool name>
Action Input: {{<json>}}
```

STRICT RULES:

- Output ONE Action, ONE Action Input, then STOP. Do NOT write a second Thought/Action in the same reply — you will only see the result of the FIRST action, so any extra ones are wasted.
- After `Action Input: {{...}}` write NOTHING more. Do not narrate, do not plan further steps.
- you will receive `Observation: {{...}}` (and any images). Read it, then send your NEXT single action.

When the evidence is sufficient, call submit_evidence.

Strategy

- You have five hypergraph gathering tools; use whichever fit the question. Call ONE per turn and read its Observation before the next.
- If the question hinges on a visual/pixel detail, open the most promising facts' rounds with get_raw_verbatim (or use search_image) so you actually SEE the images before citing.
- Be efficient: you have at most {max_steps} steps. Submit as soon as the evidence is sufficient.