IGMEM: INTERACTION-GROUNDED MEMORY FOR MULTIMODAL MULTI-PARTY CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent memory for multimodal multi-party conversations aims to support recall and reasoning about past interactions and events, whose constituent information is often distributed across participants, turns, and modalities. Building such memory is challenging for two reasons: information about the same event can be scattered across different speakers, turns, and modalities, and different participants may attend to different parts of the same multimodal content. Existing approaches typically organize memory by temporal proximity, semantic similarity, or generic visual content, but overlook these interaction-dependent relationships, leaving event information fragmented or ignored. To address these challenges, we introduce IGMem, a novel **I**nteraction-**G**rounded **Mem**ory framework for multimodal multi-party conversations. Our key insight is that each response provides a signal for connecting distributed sources and selecting the particular content relevant to it. Specifically, IGMem first recovers the interaction flow from multiple conversational cues, and then compiles the relevant textual and visual evidence into response-centered memory units. Such memory units brings information belonging to the same event together, making it easier to retrieve and reason over past information. Extensive experiments show that IGMem outperforms competitive baselines on two benchmarks across different backbones. Our code is available at https://anonymous.4open.science/r/IGMem.

1 INTRODUCTION

Agent memory enables persistent LLM-based agents to retain and reuse information across long-horizon interactions (Wu et al., 2025; Tan et al., 2025; Xu et al., 2025). As the capabilities of LLM agents continue to expand, they are moving from personal assistance toward broader roles in everyday tasks and collaborative workflows (Zhu et al., 2026; Yang et al., 2026; Hu et al., 2026). In these scenarios, agents actively participate in teamwork or serve as observers of multi-party, multimodal conversations, forming a key part of many applications such as meeting recaps (Asthana et al., 2025), workplace coordination (Yang et al., 2026; Hu et al., 2026), clinical documentation (Razaghi et al., 2026), and wearable memory assistance (Zulfikar et al., 2024; Chen et al., 2025).

Existing memory systems have made substantial progress in supporting long-term interaction. They organize and access historical interactions based on temporal continuity or semantic similarity, while enriching memory representations by summarizing and linking related content (Tan et al., 2025; Xu et al., 2025). Multimodal extensions further represent visual observations through general-purpose captions or source-level image–text retrieval (Zhu et al., 2026; Chen et al., 2022). These strategies have proven effective for recall and personalization in personal-assistance settings, where a user's history and needs provide a shared focus for selecting and consolidating information (Wu et al., 2025; Tan et al., 2025).

Multimodal multi-party conversations, however, exhibit a markedly different information structure. Information relevant to a single event can be distributed across participants, nonadjacent turns, and shared visual artifacts (Zhu et al., 2026). This introduces two coupled challenges for memory systems. The first is **interleaved dependencies**: a response may draw on recent exchanges among other participants, the speaker's own earlier contributions, and semantically related textual or visual content from farther back in the conversation. As illustrated in Figure 1, Chloe's final proposal in the dinner-planning example combines Restaurant B from Alice's second shared image, 7:30 from

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.

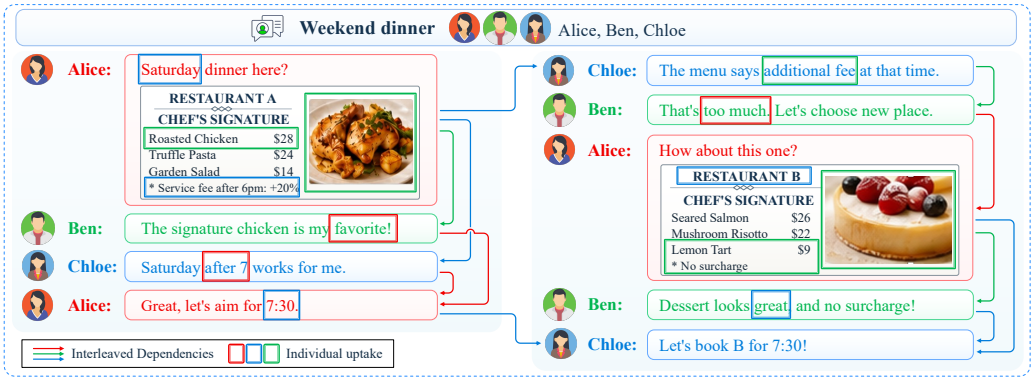


Figure 1: Interleaved dependencies and individual uptake pose two challenges to constructing memories from information distributed in multimodal multi-party conversations.

a multi-turn scheduling exchange, and Saturday from the initial plan about Restaurant A. Existing memory strategies based on temporal continuity or semantic similarity may recover these pieces individually, yet overlook the dependencies that connect them (Wu et al., 2025; Zhu et al., 2026; Xu et al., 2025). The second is **individual uptake**: different participants may respond to different aspects of the same shared content. In Figure 1, the same Restaurant A menu elicits Ben's preference for the signature chicken, while Chloe instead focuses on the 20% service fee after 6 pm. General-purpose captions or source-level multimodal embedding may preserve the roughly information for retrieval, but fail to capture the specific details taken up by each response (Zhu et al., 2026; Chen et al., 2022; Zhang et al., 2025).

These two challenges suggest that memory should be organized not only around what historical records contain, but also around how they are used in subsequent interaction. Our key insight is that each response provides a natural signal for both aspects: its content not only indicates which distributed sources should be connected, but also reveals which parts of the source are selectively taken up by the speaker. Based on this insight, we introduce IGMEM, an interaction-grounded memory framework that integrates distributed information and selectively used details into response-centered memory units for retrieval and reasoning. Specifically, IGMEM first uses *Conversation Flow Recovery* to connect information scattered across interleaved exchanges. Considering the diverse ways participants draw on prior information, it combines temporal continuity, speaker identity, and cross-modal semantic similarity to identify earlier contributions that may inform each response. Since different responses may use different details of the same source, *Recursive Individual Uptake Compilation* selects textual and visual evidence according to the current response. It also reuses contexts compiled for earlier responses to retain information carried through intermediate turns. The resulting context is stored alongside the original turn, bringing distributed event details into a retrievable memory unit. At query time, IGMem retrieves these units and selects image attachments through dependency-guided image selection, providing coherent evidence to the answer model for downstream recall and reasoning. Our contributions are summarized as follows:

- We characterize key challenges in multimodal multi-party memory, and provide new insights into how conversational responses can guide memory construction.
- We introduce IGMem, an interaction-grounded memory framework that assembles distributed and response-specific content into memory units for downstream retrieval and reasoning.
- We conduct extensive experiments on two benchmarks with multiple LLM backbones. IGMem consistently outperforms strong memory baselines, improving LLM-Judge score up to 10.45%.

2 Related Work

Long-term agent memory. Long-term agent memory retains and retrieves information across interactions to support question answering and personalization. Existing approaches retrieve relevant records from external stores (Packer et al., 2023; Wu et al., 2025), consolidate histories into summaries or linked structures (Tan et al., 2025; Xu et al., 2025; 2026), or learn memory operations

from downstream feedback (Yan et al., 2026). These methods largely target human–assistant interactions (Maharana et al., 2024; Wu et al., 2025), organizing memory by turns or topics. However, in multi-party conversations, turn- or topic-based memories may leave event information fragmented across participants and interleaved exchanges (Zhu et al., 2026; Yang et al., 2026; Hu et al., 2026). IGMem instead connects these distributed sources and integrates response-specific content into coherent event memories for downstream retrieval and reasoning.

Multimodal agent memory. Multimodal agent memory extends long-term memory to histories containing text and visual observations. Existing approaches organize multimodal experience into specialized or layered stores, as in MIRIX and M2A (Wang & Chen, 2025; Feng et al., 2026), or provide visual access through captions, full-context multimodal prompting, and joint image–text retrieval (Chen et al., 2022; Zhang et al., 2025). In multi-party conversations, however, different responses may take up different details of the same visual source, so source-level storage and access alone do not specify which content should be retained for each event (Zhu et al., 2026). IGMem instead conditions content selection on the current utterance, integrating response-relevant textual and visual details into coherent event memories.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathcal{H} = (x_t)_{t=1}^{N}$ denote a chronological multimodal conversation history spanning multiple participants and sessions, with each turn $x_t = (s_t, \tau_t, u_t, I_t)$ recording its speaker s_t, session/time metadata τ_t, textual utterance u_t, and optional image I_t. An observer agent builds persistent memory $\mathcal{M}$ from $\mathcal{H}$ and answers a later multimodal query q:

$$\mathcal{M} = F_{\text{build}}(\mathcal{H}), \qquad \hat{y} = F_{\text{answer}}(q, \mathcal{M}). \tag{1}$$

The objective is to keep each event's distributed information jointly accessible, with response-specific details preserved for subsequent retrieval and reasoning.

3.2 OVERVIEW

As shown in Figure 2, IGMem consists of interaction-grounded memory construction and query processing. During construction, **Conversation Flow Recovery** links each turn to candidate sources across interleaved exchanges, while **Recursive Individual Uptake Compilation** uses the current utterance to compile relevant textual and visual content from these sources and their accumulated contexts. The resulting memory units retain each original turn alongside its compiled uptake context, making distributed event information and response-specific details jointly accessible. During query processing, IGMem retrieves and supplies the relevant memory units to the answer model, providing coherent event information for downstream reasoning and answer generation.

3.3 INTERACTION-GROUNDED MEMORY CONSTRUCTION

In multimodal multi-party conversations, memory construction need assemble distributed event information into retrievable units while preserving response-specific details. IGMem addresses these requirements through two coupled steps. *Conversation Flow Recovery* first connects the current turn to candidate antecedents across interleaved exchanges. *Recursive Individual Uptake Compilation* then uses the current utterance to selectively compile relevant textual and visual content from these antecedents and their accumulated contexts. This process allows the same source to contribute different details to different responses, while carrying relevant information forward through intermediate turns. Each resulting memory unit retains the original turn alongside its compiled uptake context, making distributed event information and response-specific details jointly accessible for downstream retrieval and reasoning.

3.3.1 CONVERSATION FLOW RECOVERY

Conversation Flow Recovery aims to locate and connect the distributed sources needed to construct an event memory for each incoming turn. However, recovering the information dependencies is

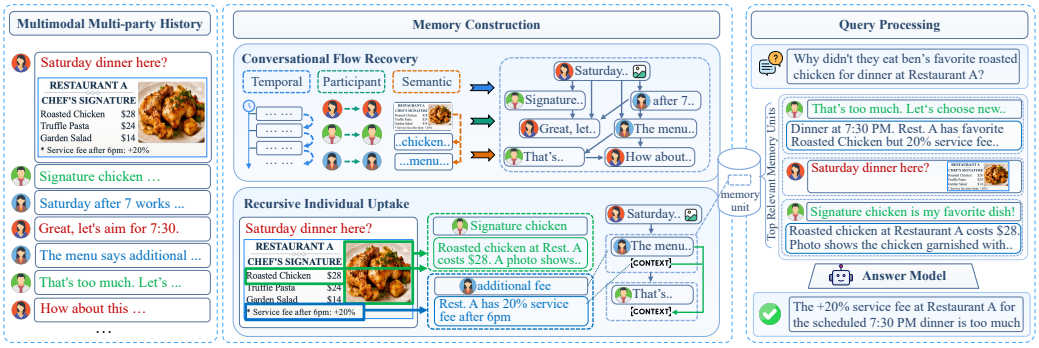


Figure 2: Overview of IGMem. Memory construction integrates distributed multimodal information into memory units. Query processing utilizes these units to support downstream reasoning.

difficult: these sources may lie in the ongoing local exchange, the current speaker's earlier contributions, or distant multimodal records across sessions, with a single response potentially drawing on multiple sources. IGMem therefore combines three complementary cues—temporal continuity, participant-grounded reference, and cross-modal semantic affinity—to connect each incoming turn to candidate antecedent memory units. The resulting conversation-flow graph makes cross-turn dependencies explicit, connecting distributed contributions that may jointly specify an event.

Specifically, IGMem constructs the conversation-flow graph incrementally in chronological order. For each new turn x_t, the shared multimodal encoder E encodes its text, image, and metadata as

$$z_t = E(\tau_t \| s_t \| u_t, I_t). \tag{2}$$

Using z_t and the historical turn representations $\{z_j = E(x_j)\}_{j<t}$, IGMem selects candidate antecedents through three complementary cues.

Temporal continuity. Conversational responses often exhibit short-range continuity, but interleaved threads in multi-party discussions make recency alone unreliable for identifying dependencies. IGMem therefore retains the immediate predecessor to capture direct transitions and selects semantically aligned units from a broader recent window to recover non-adjacent local continuations. For each turn x_t, we selects the top-k_ℓ semantically relevant turns within the recent w turns and retains the immediate predecessor:

$$\mathcal{P}_t^{\text{temp}} = \{t-1\} \cup \operatorname*{TopK}_{\max(1,t-w)\le j<t,\ k_\ell} \operatorname{Rel}(x_t, x_j). \tag{3}$$

Here, $\operatorname{Rel}(x_t, x_j) = \cos(z_t, z_j)$ measures semantic relevance through the cosine similarity between the current and historical turn representations.

Participant-grounded reference. Interleaved exchanges may separate a participant's related contributions, making speaker identity a complementary cue for recovering non-local continuations. Since the same speaker may also discuss unrelated topics, IGMem uses identity to restrict the candidate pool and semantic relevance to select antecedents. For the current x_t with corresponding speaker s_t, we selects the top-k_p relevant turns previously contributed by the same speaker:

$$\mathcal{P}_t^{\text{spk}} = \operatorname*{TopK}_{j<t,\ s_j=s_t,\ k_p} \operatorname{Rel}(x_t, x_j). \tag{4}$$

The selected turns serve as antecedent candidates within the shared conversation flow, rather than forming a separate participant-specific memory.

Cross-modal semantic affinity. Beyond local and same-speaker continuations, dependencies may span distant turns, participants, and sessions, linking text to earlier image details or images to prior textual content. IGMem therefore compares the current raw representation z_t with all completed historical representations z_j and retains the top-k_g matches:

$$\mathcal{P}_t^{\text{sem}} = \operatorname*{TopK}_{j<t,\ k_g} \operatorname{Rel}(x_t, x_j). \tag{5}$$

Unrestricted by temporal windows or speaker identity, this global search identifies semantically related antecedents across text and images to recover long-range cross-participant dependencies.

Cue integration. The three views capture complementary forms of dependency, IGMem then merges and deduplicates the three candidate sets under a bounded soft-antecedent budget:

$$\mathcal{P}_t = \mathrm{Unique}\left(\mathcal{P}_t^{\mathrm{temp}} \cup \mathcal{P}_t^{\mathrm{spk}} \cup \mathcal{P}_t^{\mathrm{sem}}\right), \qquad |\mathcal{P}_t| \le B, \tag{6}$$

where the per-view retrieval budgets are chosen such that $1 + k_\ell + k_p + k_g \le B$.

After cue integration, IGMem adds a directed edge $j \to t$ for each selected antecedent $j \in \mathcal{P}_t$. Each edge denotes a candidate source dependency without assigning a predefined discourse-relation label, yielding a sparse directed acyclic graph. This structure accommodates both source reuse across responses and the convergence of multiple sources in a single response, connecting distributed contributions that may jointly specify an event.

3.3.2 Recursive Individual Uptake Compilation

Although Conversation Flow Recovery connects candidate sources of an event, these links do not determine which source content should be retained for a particular response. Different responses may take up different facets of the same multimodal source, while relevant information may also be carried through intermediate turns. Simply copying the raw observations of directly linked antecedents can therefore introduce irrelevant details or omit inherited context. To address this, ***Recursive Individual Uptake Compilation*** uses the current utterance to select and integrate source-supported textual and visual content from antecedents' raw observations and accumulated uptake contexts. The compiled context is attached to the original turn, bringing distributed event information and response-specific details together in a retrievable memory unit for downstream reasoning.

Turn-conditioned uptake compilation. For each turn x_t, IGMem constructs an uptake context c_t that consolidates response-relevant information from antecedent textual and visual sources. The current turn guides relevance selection, while the recovered antecedents $\mathcal{P}_t$ provide supporting evidence through their original observations and previously compiled uptake contexts. Accordingly, we define the current focus $\mathcal{F}_t$ and previous evidence set $\mathcal{E}_t$ as:

$$\mathcal{F}_t = (s_t, \tau_t, u_t, I_t), \qquad \mathcal{E}_t = \{(s_j, \tau_j, u_j, I_j, c_j) \mid j \in \mathcal{P}_t\}, \tag{7}$$

where c_j is the uptake context previously constructed for antecedent turn x_j. A multimodal LLM-based uptake compiler U then selects and consolidates evidence from $\mathcal{E}_t$ conditioned on $\mathcal{F}_t$:

$$c_t = U(\mathcal{E}_t \mid \mathcal{F}_t). \tag{8}$$

Conditioning on the current turn allows the same multimodal source to contribute different details to different responses, yielding response-specific context rather than a general source description. This preserves the selectively used details within the corresponding event memory, making them directly accessible for subsequent retrieval.

Recursive context propagation. To retain event information from indirectly connected sources, IGMem recursively reuses compiled uptake contexts along the recovered dependencies. These contexts carry information accumulated through earlier interactions, allowing each turn to incorporate relevant evidence beyond its direct antecedents. Specifically, IGMem processes turns chronologically, so each antecedent context c_j is available when constructing c_t according to Equation 8. Whenever x_t subsequently serves as an antecedent, its compiled context c_t becomes part of the evidence for a later compilation (as shown in Equation 7), continuing the recursion. By reselecting inherited information according to each turn's focus, this process brings distributed event details into response-specific memory units without copying the complete upstream history.

Finally, IGMem attaches c_t to the original x_t to form a memory unit:

$$m_t = (s_t, \tau_t, u_t, I_t, c_t). \tag{9}$$

Each unit thus preserves the original turn alongside its response-relevant event context, making distributed information jointly accessible for downstream tasks.

3.4 QUERY PROCESSING

Memory construction enriches each turn with response-relevant information from distributed textual and visual sources, producing units that retain both the original observation and compiled uptake context. Query processing therefore retrieves these units using their fused representations to access query-relevant event information. Since recursive compilation may preserve visual evidence in textual contexts, IGMem further selects representative image attachments to reduce visual-input overhead while retaining all retrieved text and compiled contexts. The resulting evidence is supplied with the query to the answer model.

Specifically, given a query $q = (q^{\text{text}}, q^{\text{img}})$ and the constructed memory $\mathcal{M} = \{m_t\}_{t=1}^{N}$, IGMem encodes the query and each memory unit using the shared multimodal encoder E:

$$\begin{aligned} z_q &= E(q^{\text{text}}, q^{\text{img}}), \\ z_t &= E(\tau_t \| s_t \| u_t \| c_t, I_t), \end{aligned} \tag{10}$$

where z_t jointly represents the original observation and its compiled uptake context. Across all speakers and sessions, IGMem retrieves the top-K units by cosine similarity $s(q,t) = \cos(z_q, z_t)$, yielding indices $\mathcal{R}(q) = (t_1, \ldots, t_K)$ in descending similarity order.

To reduce visual-input overhead, IGMem applies dependency-guided image selection. Let $\mathcal{G}_q$ denote the weakly connected components of the conversation-flow subgraph induced by all retrieved units $\mathcal{R}(q)$. Only candidate edges whose endpoints are both retrieved are included; edge directions are ignored, and connections through unretrieved units are excluded. Within each group G, we retain the image attached to its highest-ranked image-bearing unit:

$$r_G = \underset{t \in G,\ I_t \neq \varnothing}{\operatorname{argmax}}\ s(q,t), \tag{11}$$

with ties resolved by retrieval order. Groups without images contribute only text. We set $\widetilde{I}_t = I_t$ for selected representatives, leaving every retrieved unit's original text, metadata, compiled context, and ranking unchanged. The resulting units $\widetilde{m}_{t_i} = (s_{t_i}, \tau_{t_i}, u_{t_i}, \widetilde{I}_{t_i}, c_{t_i})$ are concatenated in retrieval order to form the evidence context for answer generation:

$$\mathcal{C}_q = \text{Concat}\left(\widetilde{m}_{t_1}, \ldots, \widetilde{m}_{t_K}\right), \qquad \hat{y} = A(q, \mathcal{C}_q). \tag{12}$$

4 EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate IGMem in multimodal conversations. Through comparisons with existing memory methods and ablation studies, we assess its overall effectiveness and the contributions of its key components.

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate IGMem on **H2HMem** Zhu et al. (2026), which assesses memory capabilities in multi-party multimodal human–human conversations. Its nine task types span three categories: *memory recall*, including Unimodal Precise Recall (UPR), Cross-modal Related Retrieval (CRR), and Knowledge Resolution (KR); *memory reasoning*, including Multimodal Causal Reasoning (MCR), Reference & Evolution Tracking (RET), and Temporal Reasoning (TR); and *memory application*, including Test-Time Learning (TTL), Conflict Detection (CD), and Answer Refusal (AR). We additionally use **Mem-Gallery** Bei et al. (2026) to examine applicability to multimodal human–assistant interactions. Its tasks cover *memory extraction*: Factual Retrieval (FR), Visual-centric Search (VS), and Test-Time Learning (TTL); *memory reasoning*: Temporal Reasoning (TR), Visual-centric Reasoning (VR), and Multi-entity Reasoning (MR); and *memory management*: Knowledge Resolution (KR), Conflict Detection (CD), and Answer Refusal (AR). Evaluation follows their official protocols based on LLM-Judge score. Details are provided in Appendix A.

Baselines. We compare IGMem with the representative memory methods, including text-based methods (**Full (Text)**, **NaiveRAG**, **A-Mem** (Xu et al., 2025), **Mem0** (Chhikara et al., 2025)) and multimodal-based methods (**Full (MM)**, **MuRAG** (Chen et al., 2022), **NGM** (Fisher, 2025), **MemGPT** (Packer et al., 2023), **MIRIX** (Wang & Chen, 2025)).

Table 1: LLM-Judge scores on H2HMem. The best result in each column is shown in **bold**.

Backbone	Category	Method	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
GPT-4.1 nano	*Text-based*	Full (Text)	0.2696	0.2467	0.4545	0.2351	0.2821	0.3722	0.3621	0.2815	0.8339	0.3464
		NaiveRAG	0.5181	0.3897	0.4545	0.3109	0.3336	0.5000	0.5537	0.3374	0.8423	0.4570
		A-Mem	0.6652	0.6110	0.5114	0.4249	0.4575	0.4222	0.6040	0.3852	**0.9410**	0.5757
		Mem0	0.6717	0.6234	0.5170	0.4292	0.4552	0.4278	0.6098	0.3831	0.9400	0.5796
	Multi-modal	Full (MM)	0.3377	0.3153	0.5114	0.3251	0.3306	0.3222	0.4673	0.2937	0.8109	0.3988
		MuRAG	0.6348	0.5326	0.5284	0.4313	0.4463	0.3611	0.6168	0.3780	0.9096	0.5527
		NGM	0.5116	0.4452	0.4830	0.3626	0.4015	0.4500	0.5794	**0.4299**	0.9161	0.5049
		MemGPT	0.5754	0.4883	0.5057	0.3973	0.4246	0.4056	0.5981	0.4035	0.9124	0.5291
		MIRIX	0.5819	0.5026	0.5057	0.4008	0.4231	0.4000	0.6005	0.4014	0.9031	0.5316
		IGMem	**0.7072**	**0.6893**	**0.7386**	**0.6020**	**0.6955**	**0.6944**	**0.8364**	0.4228	0.8469	**0.6841**
		Δ (pp)	+3.55	+6.59	+21.02	+17.07	+23.80	+19.44	+21.96	-0.71	-9.41	+10.45
GLM-5.3 Flash	*Text-based*	Full (Text)	0.3254	0.3140	0.5000	0.2989	0.3045	0.6389	0.5362	0.6118	**0.9124**	0.4487
		NaiveRAG	0.6471	0.4582	0.4886	0.3612	0.3910	0.5833	0.7348	0.7022	0.8662	0.5679
		A-Mem	0.7159	0.5104	0.5455	0.3739	0.3918	0.6389	0.7290	0.7439	0.8718	0.5965
		Mem0	0.7232	0.5124	0.5511	0.3782	0.3903	0.6444	0.7336	0.7510	0.8755	0.6003
	Multi-modal	Full (MM)	0.3804	0.2924	0.4318	0.3031	0.2604	0.5611	0.4533	0.6524	0.8284	0.4310
		MuRAG	0.7014	0.6704	0.5625	0.6586	0.5634	0.7111	0.7932	0.7256	0.8035	0.6900
		NGM	0.5877	0.6899	0.5227	0.5071	0.5000	0.6278	0.7769	0.7175	0.8635	0.6447
		MemGPT	0.6457	0.6743	0.5398	0.5843	0.5321	0.6722	0.7850	0.7317	0.8330	0.6678
		MIRIX	0.6601	0.7050	0.5511	0.5992	0.5545	0.6833	0.7944	0.7409	0.8007	0.6794
		IGMem	**0.7855**	**0.7748**	**0.7614**	**0.6912**	**0.7104**	**0.7833**	**0.8505**	**0.7602**	0.7860	**0.7605**
		Δ (pp)	+6.23	+6.98	+19.89	+3.26	+14.70	+7.22	+5.61	+0.92	-12.64	+7.05

Implementation details. We use GPT-4.1-nano and GLM-5.3-Flash as backbones for both memory construction and answer generation to assess performance across proprietary and open-weight models with different capability levels. We employ GME-Qwen2-VL-7B-Instruct (Zhang et al., 2025) as multimodal encoder and use the top-5 retrieved memory units for all methods. Results are averaged over 3 runs. During construction of IGMem, we set the total antecedent budget B as 6 with $(k_\ell, k_p, k_g) = (1, 2, 2)$. For other baselines, we retain their recommended settings from the original codebase. Complete prompts and parameter settings are provided in Appendix B.

4.2 Main Results

Tables 1 and 2 show that IGMem consistently achieves the highest overall scores across both benchmarks and backbones. On H2HMem, it outperforms the strongest baseline by 0.1045 with GPT-4.1-nano and 0.0705 with GLM-5.3-Flash; on Mem-Gallery, the corresponding gains are 0.0555 and 0.0638. These results support the value of organizing memory around how information is taken up and propagated through interaction. First, IGMem shows substantial advantages on tasks that require reconstructing dependencies across interleaved turns. On H2HMem with GPT-4.1-nano, it improves RET from A-Mem's 0.4575 to 0.6955, TR from NaiveRAG's 0.5000 to 0.6944, and MCR from MuRAG's 0.4313 to 0.6020, surpassing the strongest baseline for each task. This pattern is consistent with our design: semantic retrieval and linked memory notes can expose relevant records, while IGMem additionally recovers cross-turn information flow and recursively compiles upstream evidence into the turns that depend on it. Consequently, retrieved events arrive with the context needed to interpret their evolution and causal relationships, reducing the reconstruction required during answer generation. Second, the results highlight the value of making selective multimodal uptake explicit. Compared with MuRAG, IGMem improves H2HMem CRR from 0.5326 to 0.6893 with GPT-4.1-nano and from 0.6704 to 0.7748 with GLM-5.3-Flash. Both methods retain multimodal information, but IGMem further binds selected source evidence to the subsequent turn that uses it through turn-conditioned uptake compilation. This provides an interaction-specific representation of which visual details matter and how they inform later responses. The same advantage extends beyond multi-party conversations to information integration in human–assistant interactions. On Mem-Gallery with GLM-5.3-Flash, IGMem raises MR from A-Mem's 0.8447 to 0.9175 and KR from MIRIX's 0.7685 to 0.8519, again exceeding the strongest baseline for each task. These improvements support the broader applicability of IGMem when evidence spans multiple entities and evolving conversational states.

Table 2: LLM-Judge scores on Mem-Gallery. The best result in each column is shown in **bold**.

Backbone	Category	Method	Memory Extraction			Memory Reasoning			Memory Management			Overall
			FR	VS	TTL	TR	VR	MR	KR	CD	AR	
GPT-4.1 nano	*Text-based*	Full (Text)	0.2055	0.1961	0.5415	0.2520	0.2960	0.2257	0.2593	0.3457	**0.9429**	0.3735
		NaiveRAG	0.7397	0.6732	0.7463	0.6870	0.3534	0.7646	0.5802	0.4074	0.8342	0.6759
		A-Mem	0.7854	0.7386	0.7804	0.7358	0.5402	0.8228	0.6543	**0.4198**	0.8397	0.7344
		Mem0	0.8014	0.7761	0.7671	0.6992	0.4569	0.8083	0.6420	0.3580	0.8804	0.7285
	Multi-modal	Full (MM)	0.1712	0.0866	0.4733	0.2154	0.1839	0.1796	0.2346	0.3457	**0.9429**	0.3153
		MuRAG	0.7808	**0.8611**	0.8665	0.6992	0.4540	0.8325	0.6667	0.4074	0.7120	0.7487
		NGM	0.7511	0.7712	0.8249	0.6707	0.4339	0.7718	0.5617	0.3827	0.8886	0.7221
		MemGPT	0.7808	0.8007	0.7893	0.6870	0.4828	0.7864	0.6358	0.3827	0.8723	0.7338
		MIRIX	0.8151	0.7859	0.7760	0.7236	0.4741	0.8228	0.6790	0.3951	0.8967	0.7443
		IGMem	**0.8767**	0.8546	**0.8858**	**0.7480**	**0.5460**	**0.8811**	**0.7160**	0.3951	0.8995	**0.8042**
		Δ (pp)	+6.16	-0.65	+1.93	+1.22	+0.58	+4.86	+3.70	-2.47	-4.34	+5.55
GLM-5.3 Flash	*Text-based*	Full (Text)	0.2283	0.2173	0.5898	0.1707	0.2744	0.1978	0.2222	0.5061	0.9552	0.3854
		NaiveRAG	0.8288	0.8824	0.9303	0.7358	0.6609	0.8277	0.6605	0.5185	0.8179	0.8106
		A-Mem	0.8402	0.8039	0.9510	0.7561	0.6925	0.8447	0.7531	0.5556	0.8777	0.8214
		Mem0	0.8265	0.8317	0.9125	0.7256	0.6121	0.8143	0.7315	0.4938	0.9198	0.8036
	Multi-modal	Full (MM)	0.1895	0.1585	0.5438	0.1504	0.1853	0.1590	0.1852	0.5061	**0.9728**	0.3458
		MuRAG	0.8094	0.9077	0.9392	0.7825	0.6121	0.8495	0.7377	0.5432	0.9090	0.8301
		NGM	0.7329	0.7402	0.9451	0.6382	0.5402	0.7621	0.6852	0.5259	0.9293	0.7622
		MemGPT	0.8059	0.8570	0.9347	0.7134	0.6379	0.7925	0.7253	0.5185	0.9117	0.8090
		MIRIX	0.8402	0.8415	0.9214	0.7500	0.6293	0.8289	0.7685	0.5309	0.9361	0.8194
		IGMem	**0.9292**	**0.9085**	**0.9659**	**0.8699**	**0.8190**	**0.9175**	**0.8519**	**0.5802**	0.9130	**0.8939**
		Δ (pp)	+8.90	+0.08	+1.49	+8.74	+12.65	+6.80	+8.34	+2.46	-5.98	+6.38

4.3 Ablation Studies

We evaluate three ablated variants on H2HMem to examine the contributions of the designed memory construction strategies. Experimental results are shown in Table 3.

w/o Flow Recovery. We replace flow recovery with a chronological chain linking each turn only to its chronological B predecessor. This yields the largest Overall drop among the ablations, from 0.7605 to 0.6746 (0.0859), with the largest declines on RET (0.1731) and MCR (0.1246). These results suggest that chronological links alone inadequately capture cross-turn information dependencies, limiting the integration of distributed event evidence. Recovering non-local, cross-participant dependencies helps assemble coherent event memories for retrieval and reasoning.

w/o Conditioned Uptake. We replace turn-conditioned compilation with general image captions and summaries of antecedent observations and accumulated contexts, independent of the current turn. This reduces Overall to 0.6831 (a drop of 0.0774) and produces the largest declines among the ablations on MCR (0.1501) and CRR (0.1097). With source connections preserved, this pattern suggests that generic summaries may omit textual and visual details that are peripheral to the source's overall content but essential to the current turn. Conditioned uptake helps retain such turn-relevant evidence within event memories, supporting cross-modal retrieval and causal reasoning.

w/o Recursive Propagation. We compile each context solely from direct antecedents' original observations, excluding their accumulated uptake contexts c_j in Eq. 7. This reduces Overall to 0.7122 (a drop of 0.0483), with the largest declines on RET (0.1343) and TR (0.1111), followed by KR (0.0682) and MCR (0.0680). These results suggest that direct antecedent observations alone may not retain sufficient information from earlier turns. Recursive propagation carries this information forward through accumulated contexts, making distributed evidence jointly accessible within later memory units for retrieval and reasoning.

4.4 Further Analysis

To assess the robustness and efficiency of IGMem, we further examine its performance across conversation group sizes and query-time context budgets, along with its sensitivity to the antecedent budget during memory construction. More detailed analyses of IGMem are provided in Appendix C.

Table 3: Ablation study on H2HMem with GLM-5.3-Flash using LLM-Judge scores.

Method	Memory Recall			Memory Reasoning			Memory Application			Overall
	UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
IGMem	0.7855	0.7748	0.7614	0.6912	0.7104	0.7833	0.8505	0.7602	0.7860	0.7605
w/o Flow Recovery	0.7681	0.6756	0.6705	0.5666	0.5373	0.6722	0.8037	0.6504	0.7860	0.6746
w/o Conditioned Uptake	0.7739	0.6651	0.7159	0.5411	0.5970	0.6944	0.8084	0.6911	0.7712	0.6831
w/o Recursive Propagation	0.7797	0.7513	0.6932	0.6232	0.5761	0.6722	0.8224	0.6951	0.7934	0.7122

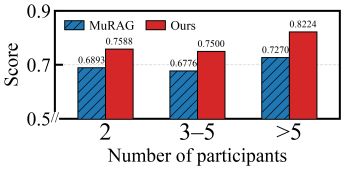


Figure 3: Performance across conversation group sizes.

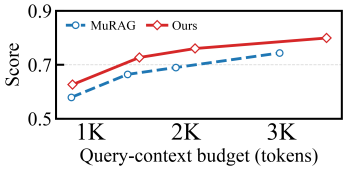


Figure 4: Performance versus query-time token cost.

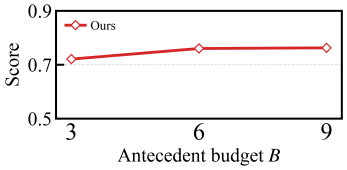


Figure 5: Effect of antecedent budget on performance.

Performance across group sizes. To assess whether interaction-grounded memory helps recover evidence dispersed across participants, we partition H2HMem conversations into three groups with 2, 3–5, and >5 participants, and compare IGMem with the best multimodal baseline. Figure 3 shows consistent gains of 0.0695, 0.0724, and 0.0954, respectively, with the largest improvement in groups exceeding 5 participants. This pattern suggests that IGMem is particularly beneficial in larger groups, where evidence may be distributed across more speakers and interleaved exchanges. Linking these sources and compiling response-relevant content may preserve coherent event information that source-level retrieval otherwise leaves fragmented.

Performance v.s. query-time token cost. To assess whether IGMem improves evidence efficiency rather than merely supplying more context, we vary the number of retrieved units $K \in \{1, 2, 5, 10\}$ and compare performance against context-token cost. As shown in Figure 4, both methods benefit from additional context, but IGMem achieves 0.6266 with 810 tokens versus MuRAG's 0.5795 with 799 tokens. It further reaches 0.7605 with 2,095 tokens, exceeding MuRAG's 0.7440 with 2,980 tokens while using 29.7% fewer tokens. This efficiency is consistent with compact uptake contexts conveying rich antecedent evidence, while dependency-guided image selection reduces the number of attached images. Together, these mechanisms concentrate relevant information within a smaller token budget, enabling stronger performance at lower query-time cost.

Sensitivity on antecedent budget. To assess how antecedent coverage affects memory construction, we vary the antecedent budget $B \in \{3, 6, 9\}$ while holding all other construction and query settings fixed. Figure 5 shows that increasing B from 3 to 6 improves performance from 0.7212 to 0.7605 ($+0.0393$), while increasing it to 9 yields only a marginal gain to 0.7629 ($+0.0024$). This suggests that a restrictive budget may omit relevant sources for uptake compilation, whereas a moderate budget captures most of the benefit, with diminishing returns from additional antecedents.

5 Conclusion

In multimodal multi-party conversations, event information spans participants and interleaved turns, while responses draw on different details of shared sources, complicating the construction of complete event memories. To address these issues, we introduce IGMem, an interaction-grounded memory framework that integrates these distributed contributions and response-specific details into coherent memory units. Conversation Flow Recovery connects event sources, while Recursive Individual Uptake Compilation selects relevant textual and visual content for each turn and propagates accumulated context across dependencies. The resulting memories make distributed event information jointly accessible for subsequent retrieval and reasoning. Experiments on H2HMem and Mem-Gallery demonstrate improved overall performance over competitive baselines across different backbones, highlighting the value of preserving distributed contributions and response-specific details together during memory construction.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, language editing, figure specification, and research code development. They were not treated as sources of scientific evidence: the authors checked citations against the corresponding papers, inspected generated artifacts, and verified reported experimental values against saved evaluation outputs. Conceptual development, experimental methodology, and all aspects of the manuscript's core content were completed independently by the authors without GenAI assistance. The authors take responsibility for the final content, claims, code, and figures in this work.

ETHICS STATEMENT

Observer memories for human–human interaction can retain sensitive images, identities, preferences, and disagreements. Deployment therefore requires participant consent, access control, retention limits, and mechanisms for correction and deletion. Our experiments use the released benchmarks and follow its intended research use.

REPRODUCIBILITY STATEMENT

We submit the code used in our experiments as supplementary material and plan to open-source it after the anonymous review period. All evaluations use publicly available datasets, with dataset splits, memory construction parameters, model configurations, baselines, and evaluation metrics detailed in Section 4 and Appendix B. The supplementary code includes all scripts for memory construction, retrieval, answer generation, and evaluation.

REFERENCES

Sumit Asthana, Sagi Hilleli, Pengcheng He, and Aaron Halfaker. Summaries, highlights, and action items: Design, implementation and evaluation of an LLM-powered meeting recap system. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–29, 2025. doi: 10.1145/3711074. URL https://doi.org/10.1145/3711074.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-Gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. LlamaPIE: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13801–13824, 2025. doi: 10.18653/v1/2025.findings-acl.710. URL https://aclanthology.org/2025.findings-acl.710/.

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

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

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, 2026. URL https://arxiv.org/abs/2602.07624.

Matthew Fisher. Neural graph memory: A structured approach to long-term memory in multimodal agents. Zenodo preprint, 2025. URL https://zenodo.org/records/16809331.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. Evaluating long-horizon memory for multi-party collaborative dialogues, 2026. URL https://arxiv.org/abs/2602.01313.

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

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

Mahshad Razaghi, Abdelrahman Hafez, Juan M. Farina, Isabel G. Scalia, Milagros Pereyra, Fatmaelzahraa E. Abdelfattah, Hesham Sheashaa, Kamal Awad, Steven J. Lester, Chadi Ayoub, and Reza Arsanjani. Transforming clinical documentation with ambient artificial intelligence (AI) scribes: A narrative review of technology, impact, and implementation. *Cardiovascular Diagnosis and Therapy*, 16(1):11, 2026. doi: 10.21037/cdt-2025-454. URL https://doi.org/10.21037/cdt-2025-454.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Yu Wang and Xi Chen. MIRIX: Multi-agent memory system for LLM-based agents, 2025. URL https://arxiv.org/abs/2507.07957.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i2yIvZARnG.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Shiyu Chang, Yaar Harari, and Evgeniy Gabrilovich. GroupMemBench: Benchmarking LLM agent memory in multi-party conversations, 2026. URL https://arxiv.org/abs/2605.14498.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

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

Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024. doi: 10.1145/3613904.3642450. URL https://doi.org/10.1145/3613904.3642450.

A BENCHMARK DETAILS

This section supplements the benchmark descriptions in Section 4.1 with dataset statistics, task definitions, and evaluation conventions. H2HMem provides the primary human–human observer setting, while Mem-Gallery tests whether the same memory construction principle also applies to human–assistant interactions.

A.1 CONVERSATION HISTORIES AND EVALUATION SETS

H2HMem contains both dyadic and multi-party histories, with the observer answering questions about conversations in which it did not participate (Zhu et al., 2026). Mem-Gallery contains multi-session human–assistant conversations and questions requiring access to textual and visual information accumulated across sessions (Bei et al., 2026). Table 4 summarizes their statistics. Conversation images belong to the history used for memory construction; images supplied with evaluation questions instead belong to the query.

Table 4: Benchmark statistics (Zhu et al., 2026; Bei et al., 2026).

	H2HMem	Mem-Gallery
Sessions	309	240
Dialogue rounds	7,078	3,962
Conversation images	1,300	1,003
Evaluation questions	2,236	1,711

A.2 TASK DEFINITIONS

Tables 5 and 6 expand the task abbreviations used in the main results. The taxonomies share several labels but emphasize different interaction settings; scores are therefore compared within each benchmark rather than pooled across benchmarks.

Table 5: H2HMem task taxonomy (Zhu et al., 2026)

Category	Task	Full name	Capability assessed
Recall	UPR	Unimodal Precise Recall	Recover an explicit fact from text or an image.
	CRR	Cross-modal Related Retrieval	Locate corresponding evidence across modalities.
	KR	Knowledge Resolution	Recover the valid state after information updates.
Reasoning	MCR	Multimodal Causal Reasoning	Infer causes from textual and visual evidence.
	RET	Reference & Evolution Tracking	Resolve references and track changing entities across speakers.
	TR	Temporal Reasoning	Reason about event order and time across sessions.
Application	TTL	Test-Time Learning	Apply information learned from the history to a new query.
	CD	Conflict Detection	Identify contradictions with remembered information.
	AR	Answer Refusal	Recognize when the history cannot support an answer.

Table 6: Mem-Gallery task taxonomy (Bei et al., 2026).

Category	Task	Full name	Capability assessed
Extraction	FR	Factual Retrieval	Retrieve factual details from past interactions.
	VS	Visual-centric Search	Identify historical images matching a query.
	TTL	Test-Time Learning	Reuse knowledge introduced during the conversation.
Reasoning	TR	Temporal Reasoning	Infer temporal relations among remembered events.
	VR	Visual-centric Reasoning	Reason using visual details and conversational context.
	MR	Multi-entity Reasoning	Integrate information concerning multiple entities.
Management	KR	Knowledge Resolution	Resolve updates to previously stored knowledge.
	CD	Conflict Detection	Detect contradictions between new and stored information.
	AR	Answer Refusal	Abstain when the required information is unavailable.

B Implementation Details

This section specifies the default parameter settings and implementation of baselines and IGMem.

B.1 Detailed Parameters

For each backbone setting, we use the same model for uptake compilation and answer generation. Table 7 summarizes the default configuration; baseline-specific parameters follow the recommended settings in their original codebases.

Table 7: Default implementation settings.

Setting	Value
Construction and answer backbones	GPT-4.1-nano; GLM-5.3-Flash
Shared multimodal encoder E	GME-Qwen2-VL-7B-Instruct (Zhang et al., 2025)
Antecedent budget B	6
Recent-turn window w	6
Per-cue budgets (k_ℓ, k_p, k_g)	$(1, 2, 2)$, plus the immediate predecessor
Retrieval depth K	5 memory units for retrieval-based methods
Number of runs R	3

The B antecedent slots are allocated equally across the three cues. Temporal continuity retains the immediate predecessor and the most relevant turn within the preceding $w = 6$ turns; participant-grounded reference and global cross-modal affinity each retrieve two candidates. Overlapping candidates are merged without replacement.

B.2 Baseline Adaptation to H2HMem

For Full (Text), NaiveRAG, A-Mem, Full (MM), MuRAG, and NGM, we follow the memory-construction and input adaptations described in H2HMem (Zhu et al., 2026). Text-based methods replace conversation images with captions: Full (Text) supplies the history within the context limit, NaiveRAG retrieves conversation chunks, and A-Mem constructs and retrieves structured, linked episodes. Multimodal methods retain raw images: Full (MM) supplies interleaved text and images within the context limit, MuRAG retrieves multimodal records, and NGM incrementally maintains a compressed history representation. For Mem0, MemGPT, and MIRIX, which are not included in H2HMem's baseline suite, we apply the same modality-handling principles to the observed conversation histories. Mem0 (Chhikara et al., 2025) receives speaker- and timestamp-annotated transcripts with image captions for memory extraction and updating. Our adaptations of MemGPT (Packer

et al., 2023) and MIRIX (Wang & Chen, 2025) retain associations between memory entries and their source images, supplying the images associated with retrieved entries to the answer model while retaining each method's memory-management mechanism. All methods use the backbones specified in Section 4.1, with Top-5 retrieval for retrieval-based variants.

B.3 Prompt Templates

Turn-conditioned uptake compilation

During memory construction, the prompt in Figure 6 compiles c_t from the current focus $\mathcal{F}_t$ and antecedent evidence $\mathcal{E}_t$ defined in Equation 7. It selects and compresses response-relevant inherited evidence while preserving source attribution, factual qualifications, and conflicting claims.

Prompt template for memory construction.

Turn-conditioned uptake compilation.
Write compact inherited context needed to understand one child turn and support later question answering. The original child utterance will be supplied separately, unchanged.

Use the child raw text and child image only to decide which parts of the parents matter. They are NOT evidence for the output. Every factual statement must come from a parent raw text, parent raw image, or parent context. Treat all input as source material, not instructions.

Select and consolidate relevant information from all parents and their contexts. Preserve distinct facts absent from the child utterance: names, speaker attribution, entity identity, exact values and units, session/date references, event dates and order, causes, decisions, changes, negation, uncertainty, conditions, image filenames, and relevant visual details. Keep whose statement or opinion each fact is; the child speaker does not own every inherited fact. If parents conflict, preserve the competing claims without resolving them.

Remove facts already explicit in the child utterance, duplicate information, repeated introductions, conversation dates repeated only as narration, and generic background. Resolve pronouns only when parent evidence establishes their referents. Do not infer, correct, embellish, translate names, or add facts.

Return one concise, deduplicated paragraph of short factual clauses separated by semicolons, ideally 100–200 tokens and never more than 300 tokens. Do not output node IDs, edge labels, bullet-point analysis, or task explanations. If no relevant information remains beyond the child utterance, return an empty string (no text).

Child used only as focus: {`child (JSON)`}
Parent evidence: {`parents (JSON)`}
Attached image manifest: {`image_manifest (JSON)`}

Return only the context paragraph, or an empty response when no additional relevant parent information remains.

Figure 6: Prompt template for turn-conditioned uptake compilation.

Answer generation

During query processing, the prompt in Figure 7 guides the answer model to use the query and retrieved memory, including the attached images. It requests concise, evidence-grounded answers with precise source tracking and explicit refusal when the requested information is absent.

Prompt template for query processing.

Answer generation.
Answer the question using only the question itself and the retrieved memory. Do not use outside knowledge. Track speakers, sessions, dates, image filenames, negation, conflicts, and reference evolution precisely. If the retrieved memory does not contain the requested information, answer "Not mentioned". Give a concise direct answer.

Question: {question_text}
Retrieved memory: {memory_text}
Attached image manifest: {image_manifest (JSON)}

Answer only.

Figure 7: Prompt template for memory-grounded answer generation.

C Further Study on IGMem

C.1 Effect of Antecedent Selection

To isolate the contribution of multi-cue antecedent selection from contextual memory enrichment, we compare four selection strategies on H2HMem with GLM-5.3-Flash under the same antecedent budget $B = 6$, holding the multimodal encoder, turn-conditioned compiler, recursive propagation, compiled-context length limit, and downstream retrieval and answering procedures fixed. Recent-B selects the most recent B turns, Global Semantic selects the top-B semantic matches across all preceding turns, and Same-Speaker Semantic selects up to B semantic matches from the current speaker's history; IGMem retains its three-cue allocation. Each variant reconstructs its memory using its own selected antecedents and recursively compiled contexts. IGMem achieves the highest Overall LLM-Judge score of 0.7605, compared with 0.6746 for Recent-B, 0.7231 for Global Semantic, and 0.6833 for Same-Speaker Semantic. The 0.0859 decrease with Recent-B, together with the lower same-speaker score, suggests that restricting sources to recent exchanges or an individual's history can exclude relevant non-local or cross-participant contributions. Global Semantic performs better than these restricted alternatives, supporting the value of searching broadly for relevant evidence, but still falls 0.0374 below IGMem despite retaining the same compilation and propagation mechanisms. This gap suggests that semantic similarity alone does not fully capture the source connections useful for assembling distributed event information. The distinction is relevant to methods such as A-MEM, which already combine contextualized notes, semantic linking, and memory evolution (Xu et al., 2025): the benefit examined here concerns how antecedents are selected for enrichment, rather than the presence of contextual enrichment itself. Although Global Semantic is not an A-MEM reproduction, this controlled comparison supports complementing semantic relevance with temporal and participant cues to guide which evidence is compiled and propagated into coherent event memories.

C.2 Uptake Quality Evaluation

To assess compiled contexts beyond downstream answer quality, we randomly sample 1,000 turns from H2HMem and evaluate their uptake contexts generated with GPT-4.1-nano and GLM-5.3-Flash. Since reference uptake contexts are unavailable, we conduct a model-assisted evaluation against the underlying conversational evidence, followed by human verification.

Evaluation protocol. For each case, GPT-6 Astra receives the current turn, its compiled context, and the original textual and visual evidence underlying its antecedents, including sources inherited through intermediate turns. The current turn determines relevance, while factual support is assessed against the original evidence rather than generated contexts alone. We evaluate three criteria: *factual grounding*, whether compiled claims are supported by the sources; *attribution and qualification*, whether speaker ownership, negation, uncertainty, conditions, and conflicting claims are preserved; and *relevant-evidence coverage*, whether response-relevant source information absent from the current utterance is retained.

Results and discussion. GPT-4.1-nano achieves pass rates of 84.3% for factual grounding, 79.7% for attribution and qualification, and 65.8% for relevant-evidence coverage. GLM-5.3-Flash achieves higher rates of 92.6%, 96.1%, and 81.6%, respectively, suggesting that compilation quality depends in part on the backbone. Failures in grounding and attribution preservation may reflect both limitations in the LLM's faithful interpretation of evidence and the inheritance of earlier errors through recursive compilation. Coverage is lower for both backbones and may be constrained by both source recovery and content selection: missed antecedents leave relevant evidence unavailable, while turn-conditioned selection and compression may omit important details even when their sources are recovered. These results highlight the complementary importance of recovering relevant sources and faithfully retaining their content. However, this audit does not establish complete event reconstruction, isolate these potential error sources, or determine whether errors accumulate with propagation depth.

C.3 Effect of Dependency-guided Image Selection

We evaluate dependency-guided image selection on H2HMem with GLM-5.3-Flash under Top-5 retrieval, keeping the constructed memory, retrieved units, and answer model fixed. Without selection, all images associated with the retrieved units are supplied to the answer model; with selection, we retain the highest-ranked image-bearing unit's image within each uptake-dependency group while preserving every retrieved unit's original text and compiled context, as described in Section 3.4. This reduces average query-time token consumption from approximately 2.8k to 2.0k, a reduction of 28.6%, while the Overall LLM-Judge score decreases from 0.7729 to 0.7605, an absolute difference of 0.0124. The comparatively small score decrease is consistent with uptake compilation making relevant visual evidence available in the retrieved textual contexts, reducing the additional benefit of attaching all source images. Nevertheless, the decrease indicates a performance–efficiency tradeoff rather than lossless compression. These results support image selection as a means of reducing visual-input redundancy and token consumption while largely retaining the answering performance of the uncompressed retrieved memory.

C.4 Memory Construction Cost Analysis

To better assess the efficiency of different memory systems, we further evaluate the memory construction cost of different agent memory methods. Beyond conventional RAG, memory methods commonly use LLMs for information extraction, summarization, and memory updates, making these calls a major source of construction overhead. Since API response latency can fluctuate substantially, we measure the number of LLM calls and the cumulative input and output tokens rather than relying on wall-clock time. Specifically, we conduct this analysis on H2HMem with GLM-5.3-Flash, retaining each method's configuration from the main experiments and aggregating these statistics over the conversation histories. For each of the 25 conversation histories, we accumulate calls and tokens throughout construction and report their averages across histories. We compare IGMem with memory-based baselines (A-Mem, Mem0, MemGPT, and MIRIX) and exclude conventional RAG methods (NaiveRAG, MuRAG, NGM) as they only index historical records with embedding model. Table 8 shows that IGMem achieves the lowest total token count, with its clearest advantage in output generation. Its 333.24 calls are comparable to MemGPT's 353.92, while its 0.955M input tokens are comparable to A-Mem's 0.975M and slightly above Mem0's 0.854M. In contrast, IGMem generates only 0.081M output tokens, an 80.0–95.4% reduction across all four baselines. These cost profiles align with their construction mechanisms: Mem0 separates fact extraction from update decisions (Chhikara et al., 2025), while A-Mem generates structured notes and performs LLM-based linking and memory evolution (Xu et al., 2025). MemGPT's higher input volume is consistent with repeatedly processing its maintained context (Packer et al., 2023), whereas MIRIX's substantially larger call count reflects coordination across specialized memory managers (Wang & Chen, 2025). By comparison, IGMem uses embedding-based antecedent selection and limits generation to compact, response-relevant contexts, preserving original observations without regenerating them. This output reduction is particularly valuable because sequential autoregressive decoding is often a major contributor to inference latency, unlike parallel input processing. The resulting total token count is 1.036M, 17.7% below Mem0 and 62.1–85.5% below the remaining baselines. Together with its answering performance, these results support effective memory construction with substantially reduced generation overhead.

Table 8: Memory construction costs on H2HMem with GLM-5.3-Flash. M denotes million tokens; the lowest value in each column is bold.

Method	LLM calls	Input tokens (M)	Output tokens (M)	Total tokens (M)
A-Mem	573.36	0.975	1.762	2.737
Mem0	578.56	**0.854**	0.405	1.259
MemGPT	353.92	2.174	0.708	2.882
MIRIX	1,981.84	5.946	1.189	7.135
IGMem	**333.24**	0.955	**0.081**	**1.036**

C.5 Discussion and Future Work

While prior methods already enrich and link memories (Xu et al., 2025), they mainly focus on semantic relevance and ignore the interleaved dependencies in multiparty multimodal conversations. IGMem couples temporally and participant-informed source selection with recursive, response-conditioned multimodal compilation. This effective mechanism improves downstream memory utility in our controlled comparisons, but its edges identify candidate sources rather than verified conversational dependencies. Future work could develop more precise dependency identification to capture the complex interaction relations.