H2HMem: A Multimodal Memory Benchmark for Agents in Human–Human Interactions

Anonymous ACL submission

Abstract

Large language model agents are increasingly deployed in human–human interaction settings, such as meeting assistants and clinical documentation systems, where they must observe conversations and retain information for downstream queries. Unlike traditional human–assistant settings, these environments are inherently multimodal, involve complex discourse phenomena such as anaphora and deixis, and contain asynchronous or conflicting information from multiple participants. However, existing memory benchmarks largely focus on single-user, text-only interactions, failing to capture these challenges. To address this gap, we introduce **H2HMem**, a **Human-to-Human Multimodal Memory Benchmark** for evaluating memory capabilities in complex human–human interactions. H2HMem includes both dyadic and multi-party conversations with multimodal information streams, and evaluates agents along three dimensions: memory recall, reasoning, and application. Experiments with advanced agents reveal substantial limitations in constructing, retaining, and utilizing memories across modalities, participants, and sessions, highlighting substantial room for improvement in next-generation LLM agents.

1 Introduction

Large language model (LLM) agents, such as ChatGPT (OpenAI, 2023) and DeepSeek (DeepSeek-AI, 2024), have advanced substantially, with memory mechanisms improving coherence in extended multi-turn human–assistant dialogues (Du, 2026). Beyond this paradigm, however, a new class of applications is emerging, as illustrated in Figure 1: **LLM agents as observers in human–human interactions**. In these settings, agents passively capture critical conversational information for subsequent querying. This capability underpins growing real-world applications, including clinical documentation systems that generate patient-centered notes from clinician–patient dialogues (Biswas and Talukdar, 2024; Vaid et al., 2026; Razaghi et al., 2026), AI-powered medical board meeting assistants processing multimodal inputs (Karthick et al., 2025), and general meeting summarization systems (Asthana et al., 2025). To operate effectively, such agents must track information distributed across multiple participants, maintain context over extended interactions, and integrate signals across modalities (Zhu and Callison-Burch, 2025). Robust multimodal memory is therefore essential.

These emerging deployment environments introduce three fundamental challenges. First, human–human conversations are inherently multimodal, naturally interleaving text with visual content such as shared photographs and screen captures (Lee et al., 2024; Guo et al., 2024). Second, natural language exhibits complex phenomena—such as anaphora and discourse deixis—that require agents to resolve references against an evolving conversational memory rather than retrieve isolated facts (Kim et al., 2022). Third, these interactions often involve multiple participants (dyadic or multi-party) who jointly shape the dialogue, contributing information asynchronously and at times presenting conflicting perspectives (Abbo et al., 2025; Liu et al., 2025). Systematically evaluating memory mechanisms under these conditions is therefore essential.

Existing memory benchmarks, however, fail to capture these complexities. Most are designed for single-user, text-only human–assistant interactions (Wu et al., 2025; Tan et al., 2025; Hu et al., 2026b; Shen et al., 2026). Although recent efforts have begun exploring human–human conversations, they remain limited in scope: LoCoMo (Maharana et al., 2024) incorporates vision but is restricted to dyadic interactions and lacks a comprehensive memory evaluation framework, whereas others (Hu et al., 2026a) support multi-party settings but remain exclusively text-based. Consequently, no

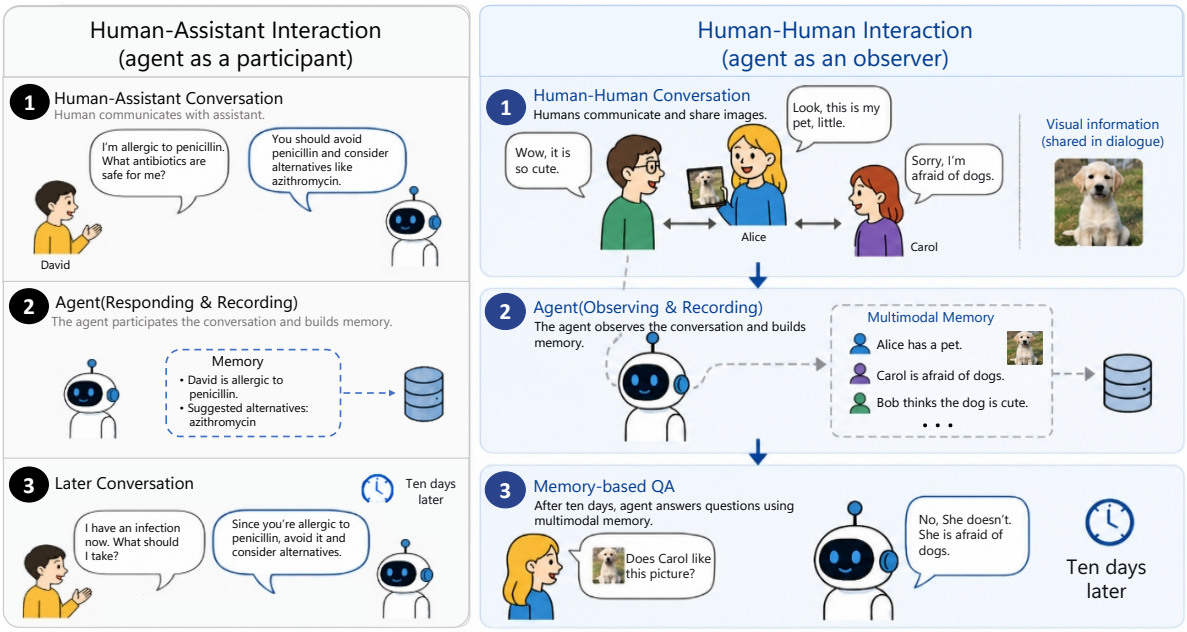

Figure 1: Comparison between Human–Assistant Interaction and Human–Human Interaction.

existing benchmark adequately captures the full spectrum of human–human interactions—spanning both dyadic and multi-party settings—while enabling multimodal memory evaluation. A comparison between existing benchmarks and ours is presented in Table 1.

To address this gap, we introduce **H2HMem**, a **Human-to-Human Multimodal Memory Benchmark**. Since directly collecting real-world multimodal conversations raises substantial privacy concerns that are difficult to fully mitigate through de-identification (Hakobyan et al., 2025; McCarthy-Howe, 2026), we develop a human-in-the-loop generation pipeline. By guiding LLM agents to iteratively generate dialogues, this pipeline avoids privacy risks while producing realistic multimodal, multi-session, and multi-participant interactions. We design evaluation tasks along three functional dimensions of memory that reflect the complexities of natural communication: (1) **Memory Recall**, which measures retrieval of multimodal facts and resolution of evolving knowledge across sessions, including Unimodal Precise Recall (UPR), Cross-modal Related Retrieval (CRR), and Knowledge Resolution (KR); (2) **Memory Reasoning**, which evaluates higher-level inference through Multimodal Causal Reasoning (MCR), Reference & Evolution Tracking (RET), and Temporal Reasoning (TR); and (3) **Memory Application**, which assesses the ability to use memory in dynamic settings through Test-Time Learning (TTL), Conflict Detection (CD), and Answer Refusal (AR). Together, these dimensions provide a comprehensive framework that moves beyond simple recall to systematically evaluate memory in complex human–human interactions. We summarize our contributions as follows:

- We introduce H2HMem, a benchmark for evaluating multimodal memory in realistic human–human observer scenarios, covering both dyadic and multi-party interactions.
- We construct a large-scale multimodal, multi-session dataset through a privacy-preserving human-in-the-loop pipeline that captures the evolving nature of real-world communication.
- We propose a comprehensive evaluation taxonomy spanning recall, reasoning, and application, revealing key limitations of current MLLMs in cross-modal memory alignment and structured reasoning.

2 Related Work

Agents in human–human Interactions. Recent work has begun to study LLM-based agents in human–human interaction settings, where the agent acts as an observer over continuous conversational streams (Asthana et al., 2025). Unlike traditional human–assistant scenarios, these settings require persistent interpretation of evolving human–human interactions and maintaining coherence over long temporal horizons (Chen et al.,

Table 1: Comparison of dialogue benchmarks. ✓: fully covered; ✗: not covered; ✗̸: partially covered. A. Round and A. Img. denote the average number of rounds and images per session, respectively, and MM-Info. indicates whether multimodal information is included.

Benchmark	Interaction Type	Conversational Characteristics			Recall			Reasoning			Application		
		A. Round	A. Img.	MM-Info.	UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR
LongMemEval (Wu et al., 2025)	Human–assistant	5.19	–	✗	✗̸	✗	✓	✗̸	✗	✓	✗	✗	✓
PersonaMem (Jiang et al., 2025)	Human–assistant	15–30	–	✗	✗̸	✗	✓	✗	✗	✓	✗	✗	✗
Mem-Gallery (Bei et al., 2026)	Human–assistant	16.51	4.18	✓	✓	✗̸	✓	✗	✗	✓	✓	✓	✓
MemoryAgentBench (Hu et al., 2026b)	Human–assistant	9.55	–	✗	✗̸	✗	✗	✗	✗	✗̸	✓	✓	✗
LoCoMo (Maharana et al., 2024)	Dyadic	10.81	3.35	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓
MSC (Xu et al., 2022)	Dyadic	8.16	–	✗	✗̸	✗	✗	✗	✗	✗	✗	✗	✗
EverMemBench (Hu et al., 2026a)	Multi-party	28.0	–	✗	✗̸	✗	✗	✗̸	✗	✓	✗	✗	✗
H2HMem	**Dyadic, Multi-party**	22.91	4.21	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2025; Vanukuru et al., 2025). As agents are deployed in increasingly rich environments, multimodal inputs—including speech, text, and documents—are incorporated (Karthick et al., 2025), significantly increasing the complexity of information flow. Commercial systems such as *Zoom AI Companion* (Zoom Video Communications, 2026) already reflect this trend, integrating multimodal meeting content for downstream querying. These characteristics jointly impose strong requirements on an agent's ability to track, integrate, and retain information over time, making memory a central capability in such settings.

Memory Mechanisms for LLM Agents. Existing memory methods for LLM agents mainly fall into three paradigms. One line of work extends the context window by directly incorporating long interaction histories into the model input (Kitaev et al., 2020; Su et al., 2024; Dao et al., 2022). Although simple, this approach incurs high computational cost, suffers from long-context degradation (Chen et al., 2023), and lacks cross-session persistence. Another line of work adopts retrieval-augmented generation, maintaining an external memory store for history retrieval (Lewis et al., 2020; Wang et al., 2023; Shi et al., 2024). While scalable and persistent, such methods mainly support factual recall and struggle with episodic dependencies and causal structures (Rajesh et al., 2026). A third direction introduces specialized memory modules with explicit operations such as writing, indexing, summarization, and forgetting (Xu et al., 2026; Kang et al., 2025). Despite these advances, existing methods are primarily developed and evaluated in human–assistant settings, leaving their effectiveness in multimodal human–human interactions unclear.

Memory Benchmarks. To evaluate memory capabilities, a range of benchmarks have been proposed. In human–assistant settings, PersonaMem (Jiang et al., 2025) studies preference following and user profiling, LongMemEval (Wu et al., 2025) focuses on long-term memory in multi-turn dialogue, Mem-Gallery (Bei et al., 2026) extends to multimodal interactions, and MemoryAgentBench (Hu et al., 2026b) evaluates memory in dialogue streams. Moving toward human–human settings, MSC (Xu et al., 2022) and LoCoMo (Maharana et al., 2024) consider conversational memory, but both are restricted to dyadic interactions. EverMemBench (Hu et al., 2026a) extends to multi-party dialogue, yet leaves multimodal aspects underexplored. Related work has also considered observer-style agents: MemBench (Tan et al., 2025) studies passive observation with one-sided inputs, and M3-Bench (Long et al., 2025) introduces video-based QA over human interactions, but is constrained by limited temporal scope. Overall, no existing benchmark jointly captures multimodality, dyadic & multi-party interaction, and long-horizon memory in human–human settings within an integrated evaluation framework. Table 1 highlights these gaps and situates our proposed H2HMem.

3 H2HMem

3.1 Problem Formulation

We study multi-session, multimodal question answering grounded in human–human interactions. An interaction, denoted as a dialogue S, is represented as a sequence of T sessions, $S = (s_1, \ldots, s_T)$. Each session s_t, associated with timestamp τ_t, corresponds to a single-day conversation, typically centered around a specific topic (e.g., movie, pet, or health).

A session is defined as $s_t = (u_{t,1}, \ldots, u_{t,n_t})$, where n_t is the number of chronologically ordered utterances. Each utterance is a multimodal tuple $u_{t,i} = (p_{t,i}, x_{t,i}, v_{t,i})$, where $p_{t,i} \in \mathcal{P}$ denotes the speaker, $x_{t,i}$ is the textual content, and $v_{t,i}$ is an

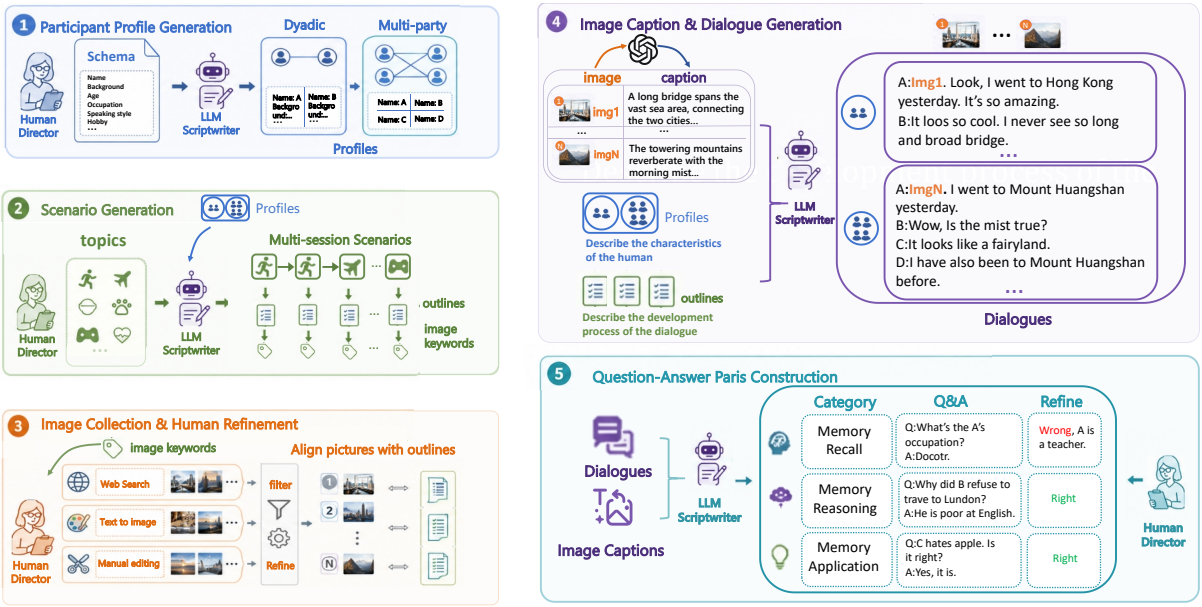

Figure 2: Dataset construction pipeline of our H2HMem. (1) generating dyadic and multi-party participant profiles from structured schemas; (2) creating multi-session scenarios with topic-specific outlines and image keywords; (3) collecting and refining images to align visual evidence with scenario outlines; (4) prompting an LLM scriptwriter with profiles, outlines, and image captions to generate dialogues; and (5) constructing and human-verifying question–answer pairs across memory recall, reasoning, and application.

optional image. The participant set $\mathcal{P}$ determines the interaction type: the dialogue is dyadic if $|\mathcal{P}| = 2$ and multi-party if $|\mathcal{P}| \geq 3$. An utterance is valid if $x_{t,i} \neq \emptyset$ or $v_{t,i} \neq \emptyset$.

We formalize memory as follows. A storage function maps each utterance $u_{t,i}$ to a memory unit m. After processing all sessions, the memory state is

$$\mathcal{M}_T = \{m_1, \ldots, m_N\}, \quad N = \sum_{k=1}^{T} n_k.$$

Given a query q, the system retrieves a subset $\mathcal{R} = \text{retrieve}(q, \mathcal{M}_T)$ and produces the final answer $a = \text{LLM}(\mathcal{R}, q)$.

Table 2: Statistics of the H2HMem dataset. "Conv. Data" denotes conversation data. "Eval. Data" denotes evaluation data. "Avg. sessions/dialogue" denotes the average number of sessions per dialogue. "Avg. rounds/session" denotes the average number of dialogue rounds per session. "Avg. participants/dialogue" denotes the average number of participants per dialogue.

H2HMem	Aspect	Dyadic	Multi-party	Sum
Conv. Data	Dialogues	20	5	25
	Sessions	284	25	309
	Dialogue Rounds	5,316	1,762	7,078
	Included Images	951	349	1,300
	Avg. participants/dialogue	2.0	5.2	2.64
	Avg. sessions/dialogue	14.2	5.0	12.4
	Avg. rounds/session	18.7	70.5	22.9
Eva. Data	QA Pairs	2,046	190	2,236
	Included Images	596	22	618

3.2 Dataset Construction Pipeline

We construct the dataset modeling human–human interactions under an online conversational setting, including dyadic and multi-party interactions. The overall statistics of H2HMem benchmark are presented in Table 2. More detailed conversation data statistics is shown in Appendix A.1. Both dialogue types follow the same pipeline with minor parameter differences. We adopt a human-in-the-loop paradigm: humans act as directors, ensuring scenario consistency, visual grounding, and quality control; LLMs serve as scriptwriters, generating dialogues, scenarios, and QA pairs. An overview of our pipeline is shown in Figure 2.

Online Conversational Setting. We focus on online conversational environments, where interactions occur via temporally ordered messages, allowing asynchronous participation, as in social media or messaging platforms (Reece et al., 2023; Zheng et al., 2022). This setting offers three key advantages: strong ecological validity, structured information flow, and support for diverse topics and participants yielding richer conversational dynamics (Wang et al., 2021; Lee and Deng, 2026).

Stage 1: Participant Profile Generation. We first define a structured schema for participant profiles, inspired by schema-guided dialogue modeling and structured persona-based datasets (Rastogi et al., 2020; Zheng et al., 2019). These profiles include

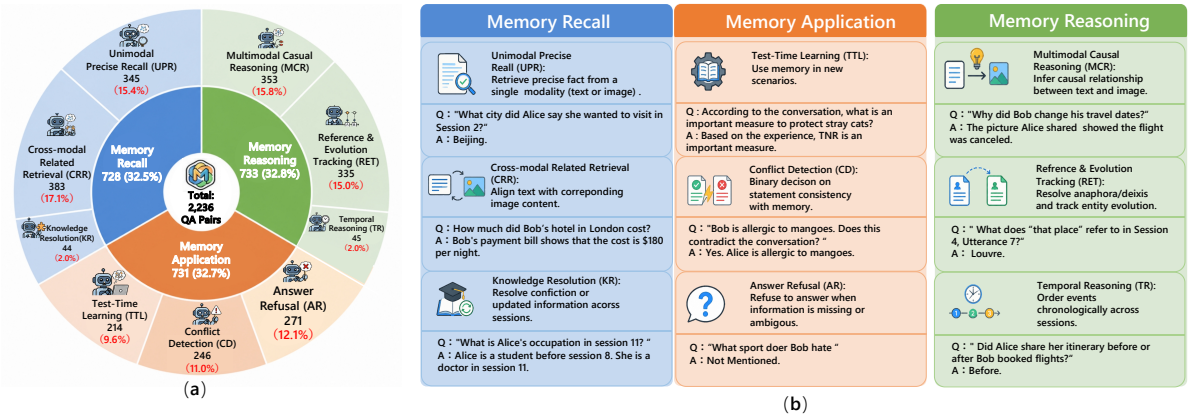

Figure 3: Figure (a) shows the total number and distribution of questions; Figure (b) provides definition and an example for each question type.

attributes such as personality, background and communication style. Conditioned on this schema, we employ DeepSeek-V3 (DeepSeek-AI, 2024) to generate structured participant profiles. An example of profiles is presented in Appendix A.2.

Stage 2: Scenario Construction. We summarize eleven common conversational topics and, given participant profiles, prompt the LLM to sample topics. For each topic, the LLM generates multiple session-level outlines, each describing a session's local events. These sessions are temporally ordered, forming a coherent multi-session scenario $S = (s_1, \ldots, s_T)$. The LLM also generates image retrieval keywords to facilitate visual content collection in the subsequent stage. An example of outlines is presented in Appendix A.3.

Stage 3: Image Collection and Human Refinement. We retrieve images through online search, supplementing retrieval with text-to-image generation (Ramesh et al., 2022; Saharia et al., 2022) and manual creation/editing based on the image keywords. Then we filter and refine pictures to align images with outlines, modifying the outlines when necessary. These images become the visual content $v_{t,i}$ in each utterance. More details are presented in Appendix A.4.1.

Stage 4: Image Captioning and Dialogue Generation. Dialogues are generated using DeepSeek-V3 (DeepSeek-AI, 2024), conditioned on participant profiles, session outlines, and images. Since DeepSeek-V3 (DeepSeek-AI, 2024) cannot process images directly, we generate detailed captions via GPT-4o (OpenAI et al., 2024b). The agent generates dialogues and refers to images using numeric identifiers. We denote each utterance as $u_{t,i} = (p_{t,i}, x_{t,i}, v_{t,i})$, where $x_{t,i}$ denotes the textual content and $v_{t,i}$ denotes the corresponding image obtained by replacing the numeric reference in $x_{t,i}$ with the actual image.

Stage 5: Question-Answer Pairs Construction. Based on the generated dialogues $S = (s_1, \ldots, s_T)$, we use DeepSeek-V3 (DeepSeek-AI, 2024) to generate a diverse set of questions q targeting different memory capabilities (recall, reasoning, application). During generation, any visual information in the dialogues is still replaced with captions. The generated question-answer pairs are further refined by human annotators to ensure clarity, correctness, and appropriate difficulty. More details on the refinement are presented in Appendix A.4.2.

3.3 Task Design

To systematically evaluate these capabilities, we design a hierarchical taxonomy of nine task types, organized into three categories. Figure 3 shows example questions for all task types.

Memory Recall. This category evaluates whether models can retrieve explicitly presented multimodal information. (1) **Unimodal Precise Recall (UPR)**: Given a query q, the model retrieves information from a single modality $x_{t,i}$ or $v_{t,i}$. (2) **Cross-modal Related Retrieval (CRR)**: The model retrieves aligned content across modalities, i.e., mapping text $x_{t,i}$ to image $v_{t,i}$ or vice versa. (3) **Knowledge Resolution (KR)**: Given multi-session dialogues $S = (s_1, \ldots, s_T)$ with updated information across sessions, the model retrieves the currently correct information from memory $\mathcal{M}_T$.

Memory Reasoning. This category evaluates reasoning over multimodal information across time and participants. (1) **Temporal Reasoning (TR)**: The model orders events across sessions using

Table 3: LLM-Judge performance with GPT-4.1-Nano (OpenAI et al., 2024a). D = Dyadic, M = Multi-party, and D&M denotes the weighted average, weighted by the number of questions. * indicates the higher value between D and M. Bold numbers in the D&M column denote the best overall performance. Light blue shading highlights D&M cells. Additional results on other backbone models are reported in Appendix D.1.

Category	Method	Dataset	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	D	0.2747*	0.2378	0.4779*	0.2422*	0.2855*	0.3929*	0.3623	0.3009*	0.8456*	0.3496*
		M	0.2155	0.3800*	0.2500	0.1413	0.2386	0.2000	0.3646*	0.1146	0.7188	0.3052
		D&M	0.2694	0.2471	0.4520	0.2351	0.2821	0.3715	0.3626	0.2820	0.8339	0.3464
	NaiveRAG	D	0.5093	0.4445	0.4896*	0.3081	0.3269	0.5000	0.5428	0.3618*	0.8467*	0.4667
		M	0.6048*	0.5104*	0.2500	0.3500*	0.4239*	0.5000	0.6400*	0.1250	0.8000	0.4933*
		D&M	0.5181	0.4489	0.4563	0.3111	0.3340	**0.5000**	0.5542	0.3377	0.8424	0.4569
	A-Mem (Xu et al., 2026)	D	0.6648	0.6070	0.5286*	0.4220	0.4515	0.4306*	0.5908	0.4014*	0.9356	0.5707
		M	0.6694*	0.6700*	0.4000	0.4600*	0.5312*	0.3500	0.7083*	0.2400	1.0000*	0.5984*
		D&M	**0.6652**	**0.6111**	0.5140	0.4247	**0.4572**	0.4216	0.6045	0.3850	**0.9415**	**0.5757**
Multi-modal	Full (MM)	D	0.3427*	0.3181*	0.5161*	0.3289*	0.3344*	0.3167	0.4681*	0.3073*	0.8107	0.4027*
		M	0.2903	0.2800	0.4500	0.2708	0.2826	0.3500*	0.4583	0.1739	0.8152*	0.3648
		D&M	0.3380	0.3156	0.5086	0.3248	0.3307	0.3204	0.4670	0.2938	0.8111	0.3988
	MuRAG (Chen et al., 2022)	D	0.6312	0.5216	0.5096	0.4407*	0.4442	0.3833*	0.6052	0.3939*	0.9002	0.5496
		M	0.6694*	0.6900*	0.6500*	0.3100	0.4762*	0.2000	0.7000*	0.2400	1.0000*	0.5757*
		D&M	0.6346	0.5326	**0.5255**	**0.4315**	0.4465	0.3629	**0.6162**	0.3782	0.9094	0.5527
	NGM (Fisher, 2025)	D	0.5119*	0.4506*	0.4872*	0.3576	0.3998	0.4562*	0.5635	0.4509*	0.9072	0.4946
		M	0.5081	0.3700	0.4500	0.4300*	0.4271*	0.4000	0.7000*	0.2400	1.0000*	0.5172*
		D&M	0.5116	0.4454	0.4830	0.3627	0.4017	0.4500	0.5794	**0.4295**	0.9157	0.5049

timestamps τ_t and utterance positions. (2) **Multimodal Causal Reasoning (MCR)**: The model infers causal relations between textual content $x_{t,i}$ and visual content $v_{t',j}$ across sessions and speakers. (3) **Reference & Evolution Tracking (RET)**: The model resolves references and tracks entity evolution across sessions s_t and speakers $p_{t,i}$.

Memory Application. This category evaluates how models apply and update memory during inference. (1) **Test-Time Learning (TTL)**: The model adapts to new scenarios at inference time by using memory $\mathcal{M}_T$. (2) **Conflict Detection (CD)**: The model detects whether a new statement contradicts $\mathcal{M}_T$. (3) **Answer Refusal (AR)**: The model refuses to answer when information is absent from $\mathcal{M}_T$ or cannot be inferred.

4 Experiment

4.1 Experimental Setup

Backbone and Memory Method. We conduct a comprehensive evaluation of both text-based and multimodal memory methods. Specifically, text-based methods include Full Memory (Text), NaiveRAG, and A-Mem (Xu et al., 2026), while multimodal memory methods include Full Memory (Multimodal), MuRAG (Chen et al., 2022), and NGM (Fisher, 2025). More detailed explanations of memory methods is shown in Appendix B. All methods are evaluated using multimodal large language models (MLLMs) as the Backbone, including the Qwen2.5-VL family (3B and 7B instruct variants) (Bai et al., 2025) and GPT-4.1-Nano (OpenAI et al., 2024a). For methods that require retrieval, we adopt a dense retriever with a default top-$K = 5$. To enable a fair comparison between text-based and multimodal memory systems, we augment textual memory methods with high-quality image captions generated by GPT-4o (OpenAI et al., 2024b). More implement details can be found in Appendix C.1.

Evaluation Metrics. To systematically assess agent performance, we adopt an LLM-as-Judge approach as our primary evaluation metric. Specifically, we employ GPT-4o-mini as a zero-shot evaluator to score each model response against the ground truth. We validated this approach by measuring agreement with human judgments on a 200-sample subset, achieving Cohen's $\kappa = 0.84$ which indicates near-perfect agreement (Landis and Koch, 1977). As a complement, we also report traditional lexical metrics, including precision, recall, F1 score, and BLEU-1. The details can be found in Appendix C.3.

4.2 Experimental Results

Table 3 and Table 4 report the overall performance evaluated by LLM-as-Judge and lexical metrics, respectively. Overall performance remains low, with the best weighted average LLM-as-Judge score reaching only 0.5757 (A-Mem (Xu et al., 2026)). Combining semantic correctness and lexical fidelity, we identify four major bottlenecks in current memory systems. **(1) Cross-modal alignment remains challenging.** A consistent gap exists between Unimodal Precise Recall (UPR) and Cross-modal Related Retrieval (CRR). For example, MuRAG (Chen et al., 2022) drops from 0.6346 to 0.5326 in LLM-as-Judge scores, with a similar lexical gap (recall: 0.4063 vs. 0.3120). **(2) Weak distractor**

Table 4: Weighted average (D&M) performance of different methods across all categories. Metrics: P=Precision, R=Recall, F1=F1-score, B=BLEU-1. Results are from GPT-4.1-nano (OpenAI et al., 2024a) with top-5 retrieval. Bold values indicate the best performance among the six methods within each metric column for the given category. Additional results on other backbone models are reported in Appendix D.2.

Category	Method	Metrics	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	P	0.1412	0.1074	0.3290	0.1197	0.1317	0.4469	0.1184	0.0650	0.8279	0.2394
		R	0.2111	0.2212	0.3120	0.1950	0.2239	0.4802	0.2260	0.0521	0.8255	0.2637
		F1	0.1479	0.1277	0.3071	0.1346	0.1470	0.3997	0.1381	0.0550	0.8215	0.2391
		B	0.1153	0.0965	0.2461	0.1069	0.1083	0.2830	0.1086	0.0489	0.8172	0.2299
	NaiveRAG	P	**0.3136**	**0.2264**	0.3249	0.1330	0.1364	0.6112	**0.1917**	**0.2378**	0.8412	**0.3082**
		R	0.3605	0.2967	0.2429	0.1843	0.1682	0.3773	0.3138	**0.2158**	0.8353	0.3042
		F1	**0.3041**	**0.2383**	0.2601	0.1420	0.1299	0.4386	0.2119	**0.2194**	0.8330	**0.2999**
		B	**0.2575**	**0.2045**	0.1656	0.1145	0.0954	0.2577	0.1754	**0.2112**	0.8309	**0.2841**
	A-Mem (Xu et al., 2026)	P	0.1384	0.0942	0.3296	0.0958	0.1103	0.2258	0.0767	0.1036	0.8834	0.2206
		R	**0.4544**	**0.4390**	**0.3988**	**0.3712**	**0.3657**	**0.6550**	**0.4325**	0.0869	**0.8979**	**0.4215**
		F1	0.1887	0.1410	**0.3483**	0.1380	0.1549	0.2895	0.1251	0.0887	0.8748	0.2364
		B	0.1257	0.0828	0.2903	0.0847	0.1020	0.2006	0.0795	0.0027	0.8690	0.2120
Multi-modal	Full (MM)	P	0.1244	0.0865	0.3053	0.1045	0.1071	0.3612	0.0968	0.1298	0.7850	0.2225
		R	0.2787	0.2558	0.3557	0.2498	0.2671	0.4891	0.2991	0.1180	0.7906	0.3034
		F1	0.1458	0.1154	0.3148	0.1343	0.1372	0.3395	0.1275	0.1175	0.7854	0.2259
		B	0.1069	0.0793	0.2575	0.0953	0.0955	0.2408	0.0843	0.1114	0.7795	0.2137
	MuRAG (Chen et al., 2022)	P	0.1747	0.0994	0.3513	0.1152	0.1299	0.3984	0.1142	0.1101	0.8856	0.2601
		R	0.4063	0.3120	0.3581	0.2923	0.3194	0.5559	0.3672	0.1067	0.8898	0.3443
		F1	0.2179	0.1349	0.3451	0.1472	0.1661	0.3657	0.1541	0.0999	0.8749	0.2738
		B	0.1529	0.0861	**0.2928**	0.1049	0.1150	0.2280	0.0951	0.0118	0.8702	0.2453
	NGM (Fisher, 2025)	P	0.2416	0.1412	**0.4053**	**0.1498**	**0.1662**	**0.6711**	0.1844	0.0780	**0.8933**	0.2858
		R	0.3471	0.2964	0.3110	0.2500	0.2692	0.5189	0.3552	0.0681	0.8862	0.3243
		F1	0.2537	0.1710	0.3438	**0.1709**	**0.1853**	**0.5255**	**0.2157**	0.0712	**0.8804**	0.2804
		B	0.1737	0.1196	0.2660	**0.1292**	**0.1393**	**0.3287**	0.1457	0.0000	**0.8747**	0.2629

filtering despite successful retrieval. A large recall–precision gap is observed across methods; for instance, A-Mem (Xu et al., 2026) achieves 0.4215 recall but only 0.2206 precision, indicating difficulty filtering noisy multiple participants' information while agents can successfully retrieve relevant history. **(3) Limited causal reasoning and adaptation to human referential conventions.** Reasoning tasks, especially Multimodal Causal Reasoning (MCR) and Reference & Evolution Tracking (RET), consistently show the lowest scores. Moreover, the near-zero BLEU-1 scores in these tasks (Table 4) indicate that models rarely reproduce the precise factual phrasing needed to connect distributed evidence, particularly under human preferences for implicit reference. **(4) Poor robustness to conflicting information.** Conflict Detection (CD) remains particularly difficult, with near-zero lexical precision and recall (e.g., A-Mem CD recall: 0.0869), highlighting the inability to resolve contradictions in human–human interactions.

Impact of Interaction Structure: Dyadic vs. Multi-party. To understand how interaction structure affects agent memory, we compare performance across dyadic and multi-party settings (Table 3). Dyadic dialogues span longer time horizons with more sessions (avg. 14.2 sessions), whereas multi-party dialogues contain denser interactions within fewer sessions (avg. 70.5 rounds/session and 5.0 sessions). This difference leads to complementary performance patterns. Consistency-oriented tasks such as Knowledge Resolution (KR) and Conflict Detection (CD) are substantially harder in multi-party settings due to contradictory signals from multiple speakers. For example, NaiveRAG's KR score drops from 0.4896 in dyadic to 0.2500 in multi-party dialogues. In contrast, tasks benefiting from concentrated contextual evidence, such as Cross-modal Related Retrieval (CRR) and Test-Time Learning (TTL), achieve comparable or higher performance in multi-party settings. Moreover, experiments with larger backbones (Qwen2.5-VL-7B-Instruct (Bai et al., 2025), detailed in Appendix D.1) show that parameter scaling alone does not eliminate this gap, indicating that current memory mechanisms remain insufficiently robust to diverse interaction structures.

Table 5: Efficiency comparison across methods. Storage is measured per session (s/sess), while retrieval and generation are measured per query (s/q).

Method	**Storage** (s/sess) ↓	**Retrieval** (s/q) ↓	**Answer** (s/q) ↓
Full (Text)	0.0015	0.1566	17.99
NaiveRAG	0.6946	1.3710	10.06
A-Mem (Xu et al., 2026)	351.08	0.0248	4.57
Full (MM)	0.0009	0.3597	26.09
MuRAG (Chen et al., 2022)	9.861	1.4674	12.64
NGM (Fisher, 2025)	6.529	0.7734	4.33

Efficiency Trade-offs. Beyond accuracy, tracking multimodal human–human interactions imposes substantial computational burdens. Table 5 reveals a clear trade-off between storage and inference latency. Full-memory methods introduce minimal storage overhead but suffer from severe inference

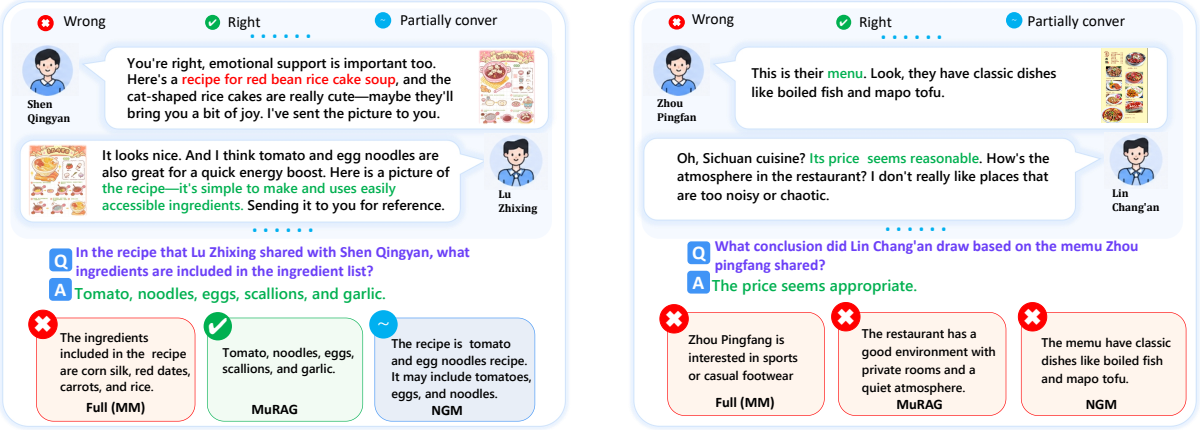

Figure 4: Case studies of multimodal conversational reasoning. (a) Identifying ingredients in Lu Zhixing's recipe. (b) Inferring Lin Chang'an's conclusion based on a shared menu.

latency, especially with multimodal inputs (17.99 s/q for Full (Text) vs. 26.09 s/q for Full (MM)). In contrast, agentic memory systems such as A-Mem (Xu et al., 2026) reduce inference latency but incur high memory construction costs (351.08 s/session). These results highlight the need for lightweight memory compression paradigms for multimodal observer agents. Additional retriever analysis is provided in Appendix D.3.

4.3 In-depth Analysis and Case Study

In-depth Analysis. To move beyond aggregated metrics, we manually analyze 100 failed cross-modal and reasoning instances from three multimodal memory methods, categorizing them into four archetypes (Table 6). The errors are highly concentrated in two major failure modes. Modal misalignment accounts for 44%–46% of cases, showing that current systems struggle to ground textual content in visual evidence. Speaker-related errors account for 32%–35%, highlighting difficulties in maintaining correct participant attribution and resolving human referential expressions in human–human interactions.

Table 6: Distribution of error types for representative multimodal methods. Error archetypes are defined as follows: *Modal Misalignment*: failing to align text with image; *Speaker-related Errors*: wrong person attribution or failing to follow human referential preference; *Temporal Confusion*: using outdated information or reversing event order; *Other / Hallucination*: remaining failure cases not covered by the above categories.

Error Archetype	Full (MM)	MuRAG (Chen et al., 2022)	NGM (Fisher, 2025)
Modal Misalignment	48%	44%	46%
Speaker-related Errors	37%	35%	32%
Temporal Confusion	15%	16%	9%
Other / Hallucination	5%	5%	6%

Case study. Two representative cases (Figure 4) illustrate how dominant failure modes manifest in practice. Case (a) focuses on ingredient identification from Lu Zhixing's recipe, requiring fine-grained visual grounding and correct speaker–image alignment; NGM (Fisher, 2025) exhibits modal misalignment by ignoring the image, while Full (MM) fails by misattributing the recipe image to the wrong speaker. Case (b) infers Lin Chang'an's conclusion drawn from a shared menu, which hinges on causal reasoning; NGM (Fisher, 2025) misattributing the conclusion to the wrong speaker, MuRAG (Chen et al., 2022) fails to align visual evidence with the textual conclusion, mistakenly presenting follow-up results instead.. Full (MM) is distracted by other visual information, causing it to provide a completely irrelevant conclusion.

5 Conclusion

We introduce H2HMem, a benchmark for evaluating multimodal memory in LLM agents within human–human interactions, providing a unified framework for assessing memory recall, reasoning, and application. Experiments show that current methods can retrieve relevant information but remain weak at integrating it. They can recall fragments — images, facts, statements — but fail to align visual evidence with text, attribute information to the correct speaker across sessions, or resolve contradictions from multiple sources. These failures persist across dyadic and multi-party settings, revealing that in multimodal human–human interactions, remembering fragments is not enough; agents must reconstruct multimodal coherent memory from distributed human communications.

Limitations

H2HMem has several limitations. First, the dataset is synthetically generated with human-in-the-loop, which may introduce artifacts or biases. Second, its scale is modest (25 dialogues, 2,236 QA pairs) and limited to English, constraining generalizability. Third, only a subset of memory methods and MLLM backbones is evaluated, leaving room for different findings with newer models. Finally, although LLM-as-Judge shows high agreement with humans, automated evaluation may still introduce bias. Despite these limitations, H2HMem provides a useful foundation for future research.

References

Giulio Antonio Abbo, Maria Jose Pinto-Bernal, Martijn Catrycke, and Tony Belpaeme. 2025. Fast multi-party open-ended conversation with a social robot. *Preprint*, arXiv:2503.15496.

Sumit Asthana, Sagi Hilleli, Pengcheng He, and Aaron Halfaker. 2025. 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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. 2026. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *Preprint*, arXiv:2601.03515.

Anjanava Biswas and Wrick Talukdar. 2024. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *International Journal of Innovative Science and Research Technology (IJISRT)*, page 994–1008.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. 2025. LlamaPIE: Proactive in-ear conversation assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13801–13824, Vienna, Austria. Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. 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*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

DeepSeek-AI. 2024. Deepseek-v3 technical report. Technical Report arXiv:2412.19437, DeepSeek-AI. Accessed: 2026-04-05.

Pengfei Du. 2026. Memory for autonomous llm agents:mechanisms, evaluation, and emerging frontiers. *Preprint*, arXiv:2603.07670.

Matthew Fisher. 2025. Neural graph memory: A structured approach to long-term memory in multimodal agents.

Peiming Guo, Sinuo Liu, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2024. An end-to-end model for photo-sharing multimodal dialogue generation. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–7.

Olya Hakobyan, Paul-Julius Hillmann, Florian Martin, Erwin Böttinger, and Hanna Drimalla. 2025. Development and evaluation of dona, a privacy-preserving donation platform for messaging data from whatsapp, facebook, and instagram. *Behavior Research Methods*, 57(3):94.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. 2026a. Evaluating long-horizon memory for multi-party collaborative dialogues. *Preprint*, arXiv:2602.01313.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2026b. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *Preprint*, arXiv:2504.14225.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25961–25970, Suzhou, China. Association for Computational Linguistics.

J. Karthick, S.S. Subithra, S. Suruthilaya, and A. Eswari. 2025. Ai-powered multimodal assistant for medical board meetings. In *2025 10th International Conference on Smart Structures and Systems (ICSSS)*, pages 1–6.

Damrin Kim, Seongsik Park, Mirae Han, and Harksoo Kim. 2022. Pipeline coreference resolution model for anaphoric identity in dialogues. In *CODI*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Meng-Chen Lee and Zhigang Deng. 2026. Multi-tpc: A multimodal dataset for three-party conversations with speech, motion, and gaze. *Scientific Data*.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2024. DialogCC: An automated pipeline for creating high-quality multimodal dialogue dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1938–1963, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zihan Liu, Parisa Rabbani, Veda Duddu, Kyle Fan, Madison Lee, and Yun Huang. 2025. The social gaze of llms: A literature review of multimodal approaches to human behavior understanding. *Preprint*, arXiv:2510.23947.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *Preprint*, arXiv:2508.09736.

Adyasha Maharana, Dong-Ho Lee, S. Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *ArXiv*, abs/2402.17753.

Thomas McCarthy-Howe. 2026. The vCon - Conversation Data Container - Overview. Internet-Draft draft-ietf-vcon-overview-01, Internet Engineering Task Force. Work in Progress.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue. Technical report, OpenAI. Accessed: 2026-04-05.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Shreyas Rajesh, Pavan Holur, Chenda Duan, David Chong, and Vwani Roychowdhury. 2026. Beyond fact retrieval: Episodic memory for rag with generative semantic workspaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32782–32790.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.

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. 2026. Transforming clinical documentation with ambient artificial intelligence (ai) scribes: a narrative review of technology, impact, and implementation. *Cardiovascular Diagnosis and Therapy*, 16(1).

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science advances*, 9(13):eadf3197.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho,

David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Ye Shen, Dun Pei, Yiqiu Guo, Junying Wang, Yijin Guo, Zicheng Zhang, Qi Jia, Jun Zhou, and Guangtao Zhai. 2026. Evolmem: A cognitive-driven benchmark for multi-session dialogue memory. *Preprint*, arXiv:2601.03543.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19336–19352, Vienna, Austria. Association for Computational Linguistics.

Samridhi Vaid, Mike Weldon, Jesse Dunn, Sacha Davis, Kevin Lonergan, Henry Li, Jeffrey Franc, Mohamed Abdalla, Daniel C. Baumgart, Jake Hayward, and J Ross Mitchell. 2026. Berta: an open-source, modular tool for ai-enabled clinical documentation. *Preprint*, arXiv:2603.23513.

Rishi Vanukuru, Payod Panda, Xinyue Chen, Ava Elizabeth Scott, Lev Tankelevitch, and Sean Rintel. 2025. Designing interfaces that support temporal work across meetings with generative ai. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, page 3600–3620, New York, NY, USA. Association for Computing Machinery.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianyi Zhang and David Traum. 2026. Rethinking evaluation in retrieval-augmented personalized dialogue: A cognitive and linguistic perspective. *Preprint*, arXiv:2603.14217.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France. European Language Resources Association.

Andrew Zhu and Chris Callison-Burch. 2025. Overhearing llm agents: A survey, taxonomy, and roadmap. *Preprint*, arXiv:2509.16325.

Zoom Video Communications. 2026. Zoom ai companion. `https://www.zoom.com/en/products/ai-assistant/`. Version 3.0, accessed April 15, 2026.

A Dataset Details

We provide more detailed dialogue statistics and additional details on key components of our data construction pipeline that are not fully elaborated in the main paper.

A.1 Conversation Data Statistics

Table 7 provides per-dialogue statistics for the 20 dyadic and 5 multi-party conversations in H2HMem. Dyadic dialogues contain 11–16 sessions per conversation (median 15) and 214–314 rounds, accompanied by 35–61 images. Multi-party dialogues are designed with fewer sessions (5 each) but much denser rounds (331–363 per dialogue) and a higher image count (69–70 per dialogue), reflecting the "intensive single-session" characteristic discussed

in the main paper. The topic coverage is also broader in dyadic settings (4–7 topics per dialogue, covering food, health, pet, shopping, travel, news, movie, entertainment, sport, and annual_summary) compared to multi-party dialogues (3–4 topics, with shopping present in all five). The time span for a single conversation is limited to one year. This detailed breakdown confirms the structural and topical diversity of our dataset, which is essential for evaluating memory under realistic human–human interaction conditions. All dialogues and question-answer pairs are in English.

A.2 Persona Schema

In the main paper, we mention that participant profiles follow a structured schema. Table 8 presents the complete schema and a concrete example from our dataset. For dyadic dialogues, we generate two participant profiles per dialogue. For multi-party dialogues, we generate four to six participant profiles per dialogue.

A.3 Outline Generation

The main paper describes that we generate session-level outlines to guide dialogue generation. Table 9 shows a complete outline example. For dyadic dialogues, we prompt the language model to sample 4–6 topics, and for each topic, generate 3–4 session-level outlines, ensuring a total of no fewer than 10 outlines per dialogue. For multi-party dialogues, we prompt the model to sample 3–4 topics, with 1–2 session-level outlines per topic, keeping the total number of outlines no more than 5 per dialogue.

A.4 Human Annotation Protocol

The main paper mentions two rounds of human verification. Below we detail the specific criteria, annotator training, and quality control procedures. Six undergraduate annotators participated, all native speakers of the dialogue language and familiar with multimodal data annotation.

A.4.1 Image Refinement

For each dialogue, the pipeline automatically retrieved or generated candidate images based on textual triggers. Six undergraduate annotators then reviewed every image together with its surrounding conversation. They focused on three aspects. First, the visual content had to match the triggering utterance exactly — for example, if a participant said "Look at this X-ray", the image had to actually show an X-ray, not a generic medical illustration.

Table 7: Statistics of dyadic and multi-party conversations in the H2HMem dataset.

Dyadic Conversations					
Dialogue	**Sessions**	**Rounds**	**Images**	**#Topics**	**Topics**
1	11	215	35	5	food, health, news, pet, shopping, sport
2	11	215	36	4	food, health, movie, pet, shopping
3	15	253	54	5	annual_summary, entertain, food, news, pet, shopping
4	15	280	61	6	entertain, food, health, movie, pet, shopping, travel
5	14	263	49	7	annual_summary, entertain, food, health, movie, news, pet, shopping
6	11	214	37	4	food, health, pet, sport, travel
7	12	221	41	4	food, health, pet, shopping, travel
8	14	254	51	6	annual_summary, food, health, movie, pet, shopping, travel
9	16	314	58	6	annual_summary, entertain, food, health, news, pet, shopping
10	16	293	44	7	annual_summary, entertain, food, health, movie, news, pet, shopping
11	13	245	36	6	annual_summary, entertain, food, health, news, pet, shopping
12	16	278	46	6	annual_summary, entertain, health, movie, news, pet, shopping
13	15	290	53	5	annual_summary, entertain, food, health, pet, shopping
14	16	297	55	6	annual_summary, entertain, health, news, pet, shopping
15	15	275	54	4	food, health, pet, shopping, sport
16	15	286	49	4	food, health, movie, pet, shopping
17	15	273	44	4	entertain, health, news, pet
18	16	310	48	4	food, health, pet, shopping, sport
19	16	308	53	4	food, health, movie, pet, shopping
20	12	232	47	5	annual_summary, food, entertain, health, pet, sport
Multi-party Conversations					
Dialogue	**Sessions**	**Rounds**	**Images**	**#Topics**	**Topics**
1	5	358	70	3	food, pet, shopping
2	5	352	70	3	sport, shopping, health
3	5	358	69	4	movie, entertain, shopping, travel
4	5	331	70	3	shopping, food, entertain
5	5	363	70	3	travel, shopping, work

Table 8: Complete participant profile schema and example.

Field	Description	Example
name	Full name	Zhao Xiaotang
age	Age	26
gender	Gender	Female
profession	Job role	Social work organization project specialist / community volunteer
title	Position	Project Supervisor
specialty	Areas of expertise	Community building, vulnerable group assistance, resource networking, public welfare project management
personality_traits	Character traits	Warm-hearted, practical and down-to-earth, strong organizational skills, slightly prone to over-worrying
core_values	Core values	Mutual assistance and sharing, community belonging, pragmatism, human warmth
fears	Anxieties or concerns	Fear of community indifference, worry about resource waste, fatigue from excessive sense of responsibility
motivations	Intrinsic drivers	Build mutual aid networks, improve community environment, help those in need, facilitate resource flow
background	Life background	From an ordinary working-class family, grew up in a company dormitory community with a strong sense of collective belonging
education	Educational background	B.A. in Social Work, East China Normal University
relationships	Social connections	Single, lives with parents, a "well-known figure" in the community, feeds three stray cats in the neighborhood

Second, the image quality needed to be sufficient for captioning and human interpretation: a resolution of at least 224×224 pixels, and no heavy blur, compression artifacts, or abstract drawings. Third, the image had to be topically appropriate and free of offensive or misleading content; in dyadic or multi-party settings, it also had to be plausible given the conversation topic (e.g., a vacation photo for travel discussions, a budget chart for a meeting). When an image failed any of these checks, annotators could replace it with a manually retrieved alternative from a stock photo library, edit the image (e.g., cropping or annotating), or — as a last resort — request a complete re-generation of the corresponding dialogue segment. The entire process of image refinement took approximately 80 person-hours.

A.4.2 QA Validation

The same annotators then examined each automatically generated question–answer pair. They checked whether the answer could be uniquely derived from the dialogue history and its images — if two different dates were mentioned without resolution, the pair was discarded or rewritten. They also ensured that the question itself was unambiguous, containing no vague references like "What about that thing?" that the context could not resolve; special attention was paid to anaphora and deixis. Finally, each QA pair was assigned to one of the nine task types (UPR, CRR, KR, MCR, RET, TR, TTL, CD, AR), and the annotators judged whether the required reasoning indeed matched the intended difficulty level (recall, reasoning, or application). Mismatched pairs were sent back for re-assignment or re-generation. The validation process took approximately 40 person-hours. Each pair was refined independently by two annotators, with cross-checking to catch individual errors.

A.4.3 Inter-Annotator Agreement

To ensure consistency, we measured Fleiss' κ for each round on a subset of 10% of the data. For image refinement, the agreement was $\kappa = 0.83$ (substantial); for QA validation, $\kappa = 0.79$ (substantial). Disagreements were resolved through discussion led by a senior researcher, who made the final decision when consensus could not be reached.

A.4.4 Time and Effort

The entire annotation process took approximately 120 person-hours: 80 hours for image refinement (including manual replacements) and 40 hours for QA validation. Each round was performed independently by two annotators, with cross-checking to catch individual errors.

A.4.5 Recruitment and Payment

Six undergraduate annotators were recruited from the authors' institution. They were compensated with a small stipend of $1 per hour, which is commensurate with local student wages.

Table 9: Example session outline.

Field	Description	Content
session_title	Brief title of the session	DIY Implementation and Neutering Plan Emergence
theme	High-level topic category	Pets - Stray Cat TNR Project
sequence_number	Order of this session in the scenario	2
timeline_date	In-story date	2024-03-15
timeline_remark	Temporal relationship to previous session	One week later, cat shelter completed
core_anchor	Key event or image trigger that drives the session	Zhao Xiaotang sends a photo of the finished foam box cat shelter (already placed, with mother cat and kittens inside) and thanks Li Yifan for his "ingenious design." She also expresses concern: "This solves the immediate crisis, but the mother cat will go into heat again soon. This can't go on indefinitely."
scenario_flow	Step-by-step description of how the dialogue should unfold	Li Yifan agrees and explains that TNR (Trap-Neuter-Return) is the most humane long-term solution internationally. He begins researching whether local animal protection organizations or pet hospitals offer discounted spay/neuter programs for strays, and creates a comparison table of costs, appointment procedures, and post-operative care. Zhao Xiaotang starts posting initiatives in the community group to see if neighbors are willing to share neutering costs or provide temporary post-surgery housing.
end_state	How the session concludes	The immediate crisis resolved, shifting focus toward a fundamental solution, beginning community mobilization and resource research.
character_states	How each participant's role or mindset evolves during the session	Zhao Xiaotang: Transitioning from hands-on rescue to systematic solution thinking, activating community organizing capabilities. Li Yifan: Shifting from technical support to solution research and cost analysis, providing decision-making basis.
key_constraints	Specific guidelines for dialogue generation (e.g., what to avoid or emphasize)	The cat shelter photo should not be praised for "looking good," but rather describe practical effects such as "the inside of the foam box is lined with aluminum insulation, with ventilation holes on top" and "the mother cat has been willing to take the kittens inside."

A.4.6 Consent

All annotators provided informed consent prior to participation. They were informed that the data would be used solely for academic research purposes.

B Baseline Models

We evaluate six baseline methods, categorized as text-based or multimodal.

B.1 Text-based Methods (with Image Captions)

For text-based memory methods, raw images are converted into image captions using GPT-4o (OpenAI et al., 2024b) before being stored in memory.

Full Memory (Text). It includes all session transcripts and image captions in textual form as part of the context, and truncates the input according to the context token limit. No retrieval or compression is applied.

NaiveRAG. It splits the conversation history into chunks. Each chunk is encoded into a semantic vector. At query time, it retrieves the top-$K = 5$ most relevant chunks based on vector similarity and concatenates them with the query.

A-Mem (Xu et al., 2026). It constructs structured memory episodes during preprocessing, where each episode summarizes a coherent discourse unit (e.g., 5–10 dialogue turns). New memories autonomously establish links to related past memories. At inference time, it retrieves relevant episodes and performs memory consolidation to support long-term reasoning.

B.2 Multimodal Methods

For multimodal memory methods, raw images are stored and retrieved directly without conversion to text.

Full Memory (Multimodal). It includes all multimodal memory information (interleaved text and raw images) as context, estimates the token consumption of images using predefined token costs, and truncates the input according to the context token limit.

MuRAG (Chen et al., 2022). It uses a dense multimodal retriever that encodes both queries and memory entries into a shared embedding space using a joint vision-language encoder. At inference time, it performs maximum inner product search over an external memory to retrieve the top-$K = 5$ most relevant multimodal passages, which are then used to augment generation.

NGM (Fisher, 2025). It proposes Neural Generative Memory, which maintains a compressed latent representation of the conversation history. Memory is updated incrementally as new dialogue turns arrive, enabling efficient long-term memorization without explicit retrieval at each step.

For methods without publicly available implementation code, we re-implement them based on the methodological descriptions provided in the original papers.

C Benchmark Evaluation Details

C.1 Implementation details

All experiments use a unified evaluation framework. Key implementation details are as follows.

Backbone MLLMs: We evaluate three vision-language models via their official APIs: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025), and GPT-4.1-Nano (OpenAI et al., 2024a). For each API call, we set the temperature to 0.1.

Retriever: Different embedding models are used for text-only and multimodal retrieval. For text-based methods, we adopt all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) as the dense retriever. For multimodal methods, we use Alibaba-NLP/gme-Qwen2-VL-7B-Instruct (Zhang et al., 2025), which jointly encodes text and images. All retrievals are performed with a default top-$K = 5$.

Image Processing. Our benchmark involves two types of images: (1) in-conversation images stored as part of the agent's memory, and (2) in-question images provided within the query during evaluation. For in-conversation images, storage format varies by memory method: text-based methods use GPT-4o (OpenAI et al., 2024b) to generate image captions (limited to 256 tokens), replacing original images with textual descriptions stored alongside dialogue transcripts, while multimodal methods resize images to 224×224 pixels and store them directly as visual tensors. During evaluation, text-based methods retrieve and reason over text-only memory entries (where images have been replaced by captions), whereas multimodal methods retrieve and utilize the original images alongside their associated text. For in-question images, all methods process them uniformly: images are resized to 224×224 pixels and fed directly to the MLLM backbone together with the textual prompt.

Memory Evaluating Prompt Template

You are a memory evaluating system. {instruction}
IMPORTANT: 1. Provide only the answer without any reasoning process. Give the answer directly in English. 2. Keep your answer within 100 words. Short and concise answers are acceptable. 3. Answer in English. This is a strict requirement. Do not answer in any other language.
{context_section}
{context_note}
Question: {question}
{format_requirement}
Examples: Question: What is the cat's name? Correct answer: Almond
Incorrect answer example (DO NOT answer like this): We need answer: cat name is Almond because...

Figure 5: Memory evaluating prompt template.

Computing Resources: All experiments were conducted on a server with four NVIDIA A10 GPUs (24GB memory each). The total computational budget across all experiments is approximately **250–300 A10 GPU hours**.

License and Terms of Use. The embedding models (all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), Alibaba-NLP/gme-Qwen2-VL-7B-Instruct (Zhang et al., 2025)) are downloaded from HuggingFace and used under their respective open-source licenses (Apache 2.0 and MIT). The Qwen2.5-VL series models (3B and 7B) (Bai et al., 2025) are accessed via Alibaba Cloud's DashScope API and used in compliance with their terms of service. The GPT series models (GPT-4o (OpenAI et al., 2024b), GPT-4.1-Nano (OpenAI et al., 2024a), GPT-4o-mini) are accessed via OpenAI's API and used in compliance with OpenAI's terms of service. The H2HMem dataset is released under a CC BY 4.0 license for research purposes only.

C.2 Evaluation Prompt Template

All models receive a unified prompt template to ensure fair comparison. The template (Figure 5) instructs the model to answer based solely on the provided conversation history (text and images) and to output the answer concisely. The temperature is fixed at 0.3 for all generations.

C.3 Evaluation Metrics

We evaluate all baseline methods from two perspectives: answer correctness and semantic alignment. All metrics are computed at the instance level and then averaged over the entire QA set.

C.3.1 LLM-as-Judge

Following prior work (Zhang and Traum, 2026), we employ GPT-4o-mini as a zero-shot evaluator. The judge prompt (provided in Figure 6) asks the model to rate each response based on semantic equivalence to the ground truth, ignoring surface-level phrasing differences.

To ensure the reliability of our primary evaluation metric, we validated the LLM-as-Judge approach against human judgments. A subset of 200 test instances was randomly sampled and independently annotated by two human evaluators. We measured the agreement between GPT-4o-mini's judgments and the human annotations using Cohen's κ, achieving a score of $\kappa = 0.84$.

C.3.2 Lexical Metrics

For tasks with deterministic ground-truth answers, we adopt four lexical overlap metrics.

Precision (P) measures the proportion of tokens in the predicted answer A_p that appear in the reference answer A_r. Let T_p and T_r denote the multisets of tokens in A_p and A_r, respectively. Precision is defined as:

$$P = \frac{|T_p \cap T_r|}{|T_p|}$$

Recall (R) measures the proportion of tokens in the reference answer A_r that are captured by the predicted answer A_p:

$$R = \frac{|T_p \cap T_r|}{|T_r|}$$

F1 Score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Judgment Prompt Template

You are an impartial judge evaluating the memory capabilities of an AI assistant with the question-answering task. Your task is to compare the Assistant's Answer against the Ground Truth and assign a score of 0, 0.25, 0.5, 0.75, or 1.
Scoring Rubric:
Score 0 (Incorrect / Miss):

- The answer contradicts the Ground Truth.
- For Yes/No questions: The answer has the wrong polarity (e.g., says "Yes" when Ground Truth is "No").
- For Open-ended questions: The answer provides factually wrong information or hallucinations.
- The assistant fails to provide the required information.

Score 0.25 (Poor / Tangential):

- The answer touches on the topic but misses the **core entity** or key value required.
- The answer contains a mix of minor correct details and **significant hallucinations** or wrong associations.
- The answer is excessively vague to the point of being useless (e.g., answering "a dog" instead of "a golden retriever").

Score 0.5 (Partial / Vague):

- The answer is technically correct, but lacks confidence or is incomplete.
- The answer captures the **main entity or concept** correctly but misses a part of the required supporting details.
- For Yes/No questions: The polarity is correct, but the reasoning is flawed (if any), or the assistant is uncertain (e.g., "I think it might be Yes").
- For Open-ended questions: The answer is too general or misses key adjectives/details present in the Ground Truth.

Score 0.75 (Good / Minor Imperfection):

- The answer is largely accurate and captures the core information confidently.
- It misses only **minor details** (e.g., specific adjectives or secondary details) that do not alter the main truth.
- The answer contains all the correct information but includes unnecessary "fluff" or slight conversational filler that reduces precision.

Score 1 (Correct / Exact):

- The answer is accurate, precise, and confident.
- For Yes/No questions: The polarity matches the Ground Truth perfectly.
- For Open-ended questions: The answer contains **all** the core information and necessary details required by the Ground Truth without hallucinations.

Input Data:

```
Question: {{question}}
Ground Truth: {{ground_truth}}
Assistant Answer: {{model_output}}
```

Output Format:
Output strictly in the following JSON format:

```
{"score": <0, 0.25, 0.5, 0.75, or 1>, "reasoning": "<short explanation>"}
```

Figure 6: Judgment prompt template for memory evaluation.

BLEU-1 measures unigram-level precision between the predicted answer A_p and the reference answer A_r. Let $\text{count}_{A_p}(w)$ denote the number of occurrences of unigram w in A_p, and $\text{count}_{A_r}(w)$ denote its occurrences in A_r. The clipped count $c(w)$ is defined as:

$$c(w) = \min\big(\text{count}_{A_p}(w), \text{count}_{A_r}(w)\big)$$

BLEU-1 is then computed as:

$$\text{BLEU-1} = \frac{\sum_{w \in A_p} c(w)}{\sum_{w \in A_p} \text{count}_{A_p}(w)}$$

D Additional Experimental Results

D.1 LLM-Judge Evaluation on Additional Backbones

In the main paper, we report LLM-Judge results using GPT-4.1-Nano (OpenAI et al., 2024a). Here, we extend the evaluation to two additional vision-language backbones: Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Table 10 and Table 11 present the results.

We highlight three key observations that directly reflect the unique characteristics of our multimodal human–human interaction benchmark:

(i) Overall performance is consistently low across backbones. Across three backbones, the absolute scores of all methods remain below 0.6 (on a 0–1 scale), with the best overall weighted average reaching only 0.5757 (A-Mem (Xu et al., 2026) on GPT-4.1-Nano (OpenAI et al., 2024a)). Even with strong memory mechanisms, models struggle to exceed this ceiling. This uniformly limited performance underscores the fundamental difficulty of our benchmark: agents must integrate information scattered across multiple participants, modalities (text and images), and sessions – a far cry from traditional single-turn or dyadic QA.

(ii) Limited benefits from model scaling for cross-modal and multi-party reasoning. Increasing parameter count from 3B to 7B yields only moderate gains, and tasks central to our benchmark — Cross-modal Related Retrieval (CRR), Multimodal Causal Reasoning (MCR), and Conflict Detection (CD) — show minimal improvement. This suggests that larger models alone cannot resolve the inherent difficulties of aligning information across participants and modalities, or of tracking evolving references across multiple sessions.

(iii) Persistent bottlenecks in reasoning over multimodal, human–human memory. Across all backbones and methods, tasks that require structured reasoning over distributed evidence — especially Multimodal Causal Reasoning (MCR), Reference & Evolution Tracking (RET), and Conflict Detection (CD) in multi-party dialogues — remain far below recall-oriented tasks. This gap confirms that current models lack robust mechanisms for maintaining coherent memory across time, speakers, and modalities, which is the central challenge posed by our human–human interaction benchmark.

These results reinforce that current systems struggle with organizing and utilizing fragmented memory, especially under the multimodal, human–human interaction captured by our benchmark. The consistency of these findings across different backbone models demonstrates that the challenges identified—cross-source alignment, structured reasoning over distributed evidence, and robustness to incomplete retrieval—are fundamental rather than artifacts of a particular model choice. Importantly, the persistence of these limitations across both proprietary and open-source backbones highlights the necessity of benchmarking conversational memory in realistic human–human interaction settings, moving beyond simplified dyadic, single-modality evaluations.

D.2 Lexical-Level Evaluation on Additional Backbones

Tables 12 and 13 report additional lexical-level results on Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), complementing the main findings presented in the paper using GPT-4.1-Nano (OpenAI et al., 2024a).

Findings are consistent across backbone models. Overall, the trends observed in the main paper remain stable across both smaller and larger Qwen backbones. In particular, lexical metrics remain uniformly low (generally below 0.4), reinforcing the conclusion that our benchmark poses a fundamentally challenging setting where exact lexical overlap is difficult due to distributed, multi-session, and multimodal information sources. This consistency suggests that the difficulty is not tied to a specific backbone, but is intrinsic to the task.

External memory consistently improves recall. Across both Qwen models, methods with external memory (e.g., NaiveRAG, MuRAG (Chen et al., 2022), A-Mem (Xu et al., 2026), NGM (Fisher, 2025)) substantially outperform full-context baselines in recall, mirroring the behavior observed with GPT-4.1-Nano (OpenAI et al., 2024a). For exam-

Table 10: LLM-Judge performance of various methods on dyadic and multi-party conversations using Qwen2.5-3B-VL-Instruct (Bai et al., 2025). D = Dyadic, M = Multi-party, D&M = Weighted average. * marks the higher value between D and M for the same method and metric. Bold numbers in D&M rows indicate the highest value among all methods for LLM-Judge performance on the weighted average dataset. Light blue background highlights the D&M data cells.

Category	Method	Dataset	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	D	0.2369*	0.2043*	0.3929*	0.2338*	0.2171	0.3846*	0.3463	0.2489*	0.9479	0.3284
		M	0.3145	0.2600	0.3500	0.2000	0.3542*	0.3000	0.3900*	0.1200	0.9783*	0.3630*
		D&M	0.2439	0.2080	0.3880	0.2314	0.2269	0.3752	0.3514	0.2358	0.9507	0.3313
	NaiveRAG	D	0.5185	0.4207	0.4392	0.2855	0.3189	0.3500	0.5241	0.3341*	0.8957*	0.4667
		M	0.6048*	0.6700*	0.4500*	0.3300*	0.4130*	0.4500*	0.5900*	0.2000	0.8229	0.5173*
		D&M	0.5263	0.4370	0.4404	0.2887	0.3257	0.3611	0.5318	0.3204	0.8890	0.4607
	A-Mem (Xu et al., 2026)	D	0.6786*	0.5453	0.4872*	0.3735	0.3349	0.4125*	0.6053	0.3664*	0.9258	0.5273
		M	0.6774	0.6500*	0.3000	0.3200	0.3854*	0.4000	0.6300*	0.2400	1.0000*	0.5465*
		D&M	**0.6785**	**0.5521**	0.4659	0.3697	0.3385	**0.4111**	**0.6082**	0.3536	0.9326	**0.5292**
Multi-modal	Full (MM)	D	0.3493*	0.2783*	0.5128*	0.2384*	0.2913	0.3654	0.4535	0.2283*	0.9516*	0.3829*
		M	0.3250	0.2188	0.3500	0.2188	0.4062*	0.4000*	0.5300*	0.0600	0.9375	0.3817
		D&M	0.3471	0.2744	0.4943	0.2370	0.2995	0.3692	0.4624	0.2112	0.9503	0.3827
	MuRAG (Chen et al., 2022)	D	0.6115	0.5348	0.4539*	0.3927	0.3604	0.4062*	0.5865	0.3608*	0.9128	0.5236
		M	0.6210*	0.6500*	0.4500	0.4600*	0.4062*	0.2000	0.6300*	0.1600	1.0000*	0.5489*
		D&M	0.6124	0.5423	0.4535	0.3974	0.3637	0.3833	0.5916	0.3405	0.9208	0.5257
	NGM (Fisher, 2025)	D	0.4738	0.4351*	0.5128*	0.3182	0.3213	0.3250*	0.5548	0.4206*	0.9645	0.4671
		M	0.4758*	0.4200	0.4000	0.3100	0.3750*	0.3000	0.7300*	0.1200	1.0000*	0.4802*
		D&M	0.4740	0.4341	0.5000	0.3176	0.3251	0.3222	0.5752	0.3902	0.9678	0.4682

Table 11: LLM-Judge performance of various methods on dyadic and multi-party conversations using Qwen2.5-VL-7B-Instruct (Bai et al., 2025). D = Dyadic, M = Multi-party, D&M = Weighted average. * marks the higher value between D and M for the same method and metric. Bold numbers in D&M rows indicate the highest value among all methods for LLM-Judge performance on the weighted average dataset. Light blue background highlights the D&M data cells.

Category	Method	Dataset	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	D	0.2667*	0.2352	0.3716*	0.2389*	0.2632*	0.4392*	0.3878	0.2661*	0.8817	0.3480*
		M	0.1897	0.3646*	0.2500	0.1800	0.2500	0.2000	0.3900*	0.1562	0.9457*	0.3388
		D&M	0.2630	0.2408	0.3650	0.2356	0.2625	0.4227	0.3880	0.2589	0.8861	0.3474
	NaiveRAG	D	0.4918	0.4096	0.5312*	0.1870	0.2795	0.2143	0.5156	0.2066*	0.8643	0.4127
		M	0.5104*	0.3889	0.1000	0.2083*	0.1719	0.2500*	0.6806*	0.1389	0.9444*	0.4232*
		D&M	0.4937	0.4067	0.4741	0.1891	0.2685	0.2183	0.5372	0.1990	0.8733	0.4135
	A-Mem (Xu et al., 2026)	D	0.6330	0.5524	0.5577*	0.4126	0.3955	0.5000*	0.6144	0.3909*	0.8860	0.5429
		M	0.6048	0.6700*	0.3500	0.5100*	0.4167*	0.4000	0.5833	0.0800	0.8854	0.5279
		D&M	**0.6306**	**0.5585**	0.5405	**0.4201**	**0.3977**	**0.4914**	**0.6113**	0.3656	0.8860	**0.5415**
Multi-modal	Full (MM)	D	0.3681*	0.2514*	0.5203*	0.2640*	0.2782	0.3875*	0.3889	0.1714*	0.7820	0.3512*
		M	0.3190	0.2083	0.4500	0.2292	0.2727*	0.3500	0.3854	0.1042	0.8478*	0.3389
		D&M	0.3651	0.2488	0.5148	0.2617	0.2779	0.3844	0.3884	0.1660	0.7880	0.3503
	MuRAG (Chen et al., 2022)	D	0.5924	0.5271	0.5705*	0.3843	0.3640	0.4437*	0.5726	0.3553*	0.7998	0.5053
		M	0.5645	0.6700*	0.4500	0.4700*	0.3750*	0.2000	0.6400*	0.1250	1.0000*	0.5361*
		D&M	0.5899	0.5356	**0.5579**	0.3904	0.3652	0.4220	0.5818	0.3342	0.8196	0.5079
	NGM (Fisher, 2025)	D	0.4976	0.4630*	0.4936*	0.3182	0.3798*	0.4000*	0.6000	0.3779*	0.8721	0.4782
		M	0.4758	0.4200	0.4000	0.3100	0.3750	0.3000	0.7300*	0.1200	1.0000*	0.4802*
		D&M	0.4956	0.4591	0.4864	0.3174	0.3794	0.3927	0.6160	**0.3567**	**0.8847**	0.4783

Table 12: Weighted average (D&M) performance of different methods (Qwen2.5-3B-VL-Instruct (Bai et al., 2025), top-5 retrieval) across categories. Metrics: P=Precision, R=Recall, F1=F1-score, B=BLEU-1. Bold values indicate the best performance among the six methods within each metric column for the given category.

Category	Method	Metrics	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	P	0.1274	0.0830	0.3095	0.0950	0.1486	0.3775	0.1334	0.0448	**0.9433**	0.2293
		R	0.2236	0.1922	0.2966	0.2074	0.1788	0.4185	0.1977	0.0360	**0.9361**	0.2548
		F1	0.1428	0.1008	0.2862	0.1113	0.1348	0.3187	0.1337	0.0372	**0.9344**	0.2295
		B	0.1049	0.0736	0.2159	0.0807	0.0887	**0.2038**	0.1096	0.0340	**0.9322**	0.2206
	NaiveRAG	P	**0.2148**	**0.1686**	**0.4127**	**0.1164**	**0.1550**	**0.4592**	**0.1693**	**0.1932**	0.8733	**0.2992**
		R	0.3776	0.3079	0.3066	0.1959	0.2368	0.4288	0.3582	**0.1694**	0.8670	0.2957
		F1	**0.2390**	**0.1888**	0.3383	0.1283	0.1649	**0.3596**	**0.1960**	**0.1734**	0.8662	**0.2894**
		B	**0.1888**	**0.1474**	0.2330	**0.0976**	**0.1220**	0.2003	**0.1481**	**0.1663**	0.8640	**0.2743**
	A-Mem (Xu et al., 2026)	P	0.1599	0.1120	0.3608	0.1027	0.1342	0.2264	0.1071	0.0987	0.8959	0.2051
		R	**0.4590**	**0.4143**	**0.4391**	**0.3279**	**0.3326**	**0.4899**	**0.4297**	0.0710	0.9120	**0.4160**
		F1	0.2134	0.1622	**0.3832**	**0.1412**	**0.1694**	0.2698	0.1568	0.0733	0.8915	0.2257
		B	0.1496	0.1000	**0.3241**	0.0931	0.1195	0.1915	0.1026	0.0016	0.8848	0.2038
Multi-modal	Full (MM)	P	0.1681	0.0891	0.3617	0.1077	0.1481	**0.5761**	0.1261	0.0492	**0.9390**	0.2196
		R	0.2329	0.1516	0.3822	0.1610	0.2295	0.4177	0.2593	0.0434	**0.9323**	0.2967
		F1	0.1703	0.0979	0.3540	0.1160	0.1493	0.4057	0.1346	0.0427	**0.9303**	0.2194
		B	0.1394	0.0774	0.2809	0.0929	0.1042	0.2239	0.0972	**0.0418**	**0.9286**	0.2086
	MuRAG (Chen et al., 2022)	P	0.1959	0.1272	0.3653	0.1365	0.1449	0.3337	0.1425	**0.0729**	0.9323	0.2424
		R	**0.4253**	**0.3320**	**0.3862**	**0.2932**	**0.3220**	**0.5310**	0.3537	**0.0539**	0.9241	**0.3371**
		F1	0.2349	0.1661	**0.3653**	0.1693	0.1749	0.3426	0.1735	**0.0556**	0.9219	0.2666
		B	0.1528	0.1109	**0.3081**	0.1229	0.1252	0.2357	0.1061	0.0000	0.9149	0.2369
	NGM (Fisher, 2025)	P	**0.3179**	**0.1627**	**0.3839**	**0.1579**	**0.1805**	0.5122	**0.2573**	0.0490	0.9365	**0.2866**
		R	0.3763	0.2830	0.3196	0.2575	0.2556	0.4434	**0.3672**	0.0366	0.9311	0.3244
		F1	**0.3107**	**0.1845**	0.3345	**0.1736**	**0.1867**	**0.4068**	**0.2738**	0.0369	0.9284	**0.2792**
		B	**0.2011**	**0.1213**	0.2489	**0.1267**	**0.1310**	**0.2473**	**0.1768**	0.0000	0.9249	**0.2607**

Table 13: Weighted average (D&M) performance of different methods (Qwen2.5-7B-VL-Instruct (Bai et al., 2025), top-5 retrieval) across categories. Metrics: P=Precision, R=Recall, F1=F1-score, B=BLEU-1. Bold values indicate the best performance among the six methods within each metric column for the given category.

Category	Method	Metrics	Memory Recall			Memory Reasoning			Memory Application			Overall
			UPR	CRR	KR	MCR	RET	TR	TTL	CD	AR	
Text-based	Full (Text)	P	0.0784	0.0599	0.1913	0.0712	0.0764	0.3516	0.0826	0.0528	**0.8654**	0.1745
		R	0.2014	0.2270	0.3362	0.2186	0.2332	0.4880	0.2409	0.0393	**0.8574**	0.2871
		F1	0.0984	0.0822	0.2358	0.0945	0.1014	0.3273	0.1023	0.0410	**0.8575**	0.1888
		B	0.0699	0.0529	0.1988	0.0668	0.0724	0.2279	0.0709	0.0372	**0.8535**	0.1641
	NaiveRAG	P	**0.2876**	**0.1741**	**0.2779**	0.0837	**0.1283**	0.1629	**0.1536**	**0.0716**	0.8571	0.2560
		R	0.2999	0.1952	0.1328	0.0868	0.1151	0.2654	0.2208	**0.0599**	0.8515	0.2819
		F1	**0.2624**	**0.1663**	0.1666	0.0731	0.0990	0.1288	**0.1629**	**0.0607**	0.8500	0.2569
		B	**0.2177**	**0.1325**	0.0605	0.0533	0.0558	0.0752	**0.1298**	**0.0573**	0.8476	0.2415
	A-Mem (Xu et al., 2026)	P	0.0739	0.0500	0.1962	0.0627	0.0624	0.0943	0.0506	0.0545	0.5624	0.1894
		R	**0.4381**	**0.4206**	**0.3834**	**0.3074**	**0.3397**	0.5892	**0.3428**	0.0427	0.8198	**0.4107**
		F1	0.1185	0.0822	0.2519	0.0939	0.0982	0.1465	0.0826	0.0425	0.5707	0.1943
		B	0.0972	0.0385	0.2024	0.0586	0.0680	0.0955	0.0569	0.0010	0.5555	0.1702
Multi-modal	Full (MM)	P	0.0971	0.0470	0.1770	0.0494	0.0524	0.2606	0.0477	0.0535	0.7640	0.2047
		R	0.2361	0.1862	0.3422	0.1761	0.2606	0.5056	0.2412	0.0411	0.7578	0.2936
		F1	0.1171	0.0633	0.2244	0.0683	0.0763	0.2210	0.0677	0.0426	0.7546	0.2070
		B	0.0894	0.0420	0.1810	0.0474	0.0529	0.1208	0.0430	0.0394	0.7431	0.1935
	MuRAG (Chen et al., 2022)	P	0.1247	0.0716	0.2083	0.0929	0.0745	0.1945	0.0756	0.0554	0.7862	0.2286
		R	0.3999	0.3402	0.3413	0.2774	0.3070	**0.5959**	0.3282	0.0468	0.8220	0.3330
		F1	0.1692	0.1032	0.2518	0.1241	0.1069	0.2249	0.1046	0.0483	0.7805	0.2354
		B	0.1279	0.0607	0.2055	0.0872	0.0781	0.1524	0.0617	0.0020	0.7664	0.2089
	NGM (Fisher, 2025)	P	0.1925	0.0884	0.2652	**0.1070**	0.0981	**0.4642**	0.1246	0.0605	0.8435	**0.2732**
		R	0.3505	0.2925	0.3155	0.2393	0.2755	0.5359	0.3089	0.0487	0.8461	0.3213
		F1	0.2185	0.1149	**0.2784**	**0.1372**	**0.1262**	**0.4224**	0.1588	0.0497	0.8391	**0.2678**
		B	0.1478	0.0760	**0.2223**	**0.0978**	**0.0989**	**0.2751**	0.1001	0.0000	0.8276	**0.2427**

ple, A-Mem (Xu et al., 2026) achieves the highest recall across most settings (e.g., 0.4160 on Qwen2.5-VL-3B-Instruct (Bai et al., 2025) and 0.4107 on Qwen2.5-VL-3B-Instruct (Bai et al., 2025)), indicating that structured memory access is critical for recovering dispersed information. However, precision remains low across all methods, confirming that retrieved evidence often contains noise.

Cross-modal retrieval remains a bottleneck. The gap between unimodal recall (UPR) and cross-modal retrieval (CRR) persists across both Qwen backbones. Even multimodal retrieval methods such as MuRAG (Chen et al., 2022) and NGM (Fisher, 2025) fail to close this gap, highlighting the continued difficulty of aligning textual queries with visual content in human–human interaction settings.

Reasoning-intensive tasks show extremely low lexical overlap. Tasks such as MCR, RET, and especially CD continue to exhibit near-zero BLEU-1 scores across both models. This further confirms that lexical metrics are poorly suited for evaluating open-ended reasoning and decision-making tasks, and that models struggle to produce lexically aligned outputs even when reasoning is partially correct.

Model scaling provides limited gains. Comparing Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), improvements from scaling are modest and inconsistent. While the 7B model shows slight gains in some recall metrics, it does not fundamentally change the performance landscape. This observation aligns with the main paper: increasing model capacity alone is insufficient to address challenges in memory retrieval, cross-modal grounding, and multi-session reasoning.

D.3 Retriever Analysis

Table 14 reports the performance under different top-K retrieval settings. We observe that increasing K generally improves performance for most methods, suggesting that retrieving more candidate memories helps alleviate missing relevant information. However, the gain is not monotonic. For example, A-Mem (Xu et al., 2026) and NGM (Fisher, 2025) both achieve their best performance at $K = 15$, after which performance slightly declines. This indicates that introducing excessive retrieved content may introduce noise and negatively affect downstream reasoning. Different methods also exhibit varying sensitivity to K. NaiveRAG shows a relatively steady improvement as K increases, implying that its simple retrieval-and-generate pipeline primarily benefits from increased recall. In contrast, MuRAG (Chen et al., 2022) reaches its peak at $K = 10$ and then degrades, suggesting that multimodal retrieval may be more susceptible to noise accumulation when irrelevant information is introduced. Overall, these results highlight a trade-off between recall and noise: while larger K improves coverage, it may also harm performance due to irrelevant or redundant information. This underscores the importance of effective retrieval filtering and ranking strategies, rather than simply increasing the number of retrieved candidates.

Table 14: Performance under different top-K settings. The best result in each row is highlighted in bold.

Method	top-5	top-10	top-15	top-20
NaiveRAG	0.4569	0.5023	0.5401	**0.5449**
A-Mem (Xu et al., 2026)	0.5757	0.6277	**0.6428**	0.6380
MuRAG (Chen et al., 2022)	0.5527	**0.5902**	0.5665	0.5726
NGM (Fisher, 2025)	0.5049	0.6253	**0.6277**	0.6213

E Potential Risks

The main risks of this work include potential misuse of the benchmark for surveillance applications and overgeneralization of results beyond English and synthetic data. The H2HMem dataset is released under a CC BY license to encourage broad use and reproducibility, but we urge researchers to deploy the benchmark only for benign applications such as meeting assistants and clinical documentation systems.

F Use of AI Assistants

LLMs are used in this work strictly as auxiliary tools for data construction and evaluation. In dataset construction , LLMs assist in generating participant profiles, multi-session scenario outlines, dialogue scripts, and candidate question–answer pairs. To address the limitation that LLMs cannot process images directly, they are also used to produce detailed captions for all images in the dialogues. All LLM-generated content is subsequently reviewed, refined, and verified by human annotators to ensure quality and correctness. In our evaluation, we adopt an LLM-as-Judge to score model responses against ground truth answers, following common practice in prior work. All technical content, experimental analysis, and scientific claims are authored and finalized by the authors.