BEYOND DYADIC MEMORY: INTERACTION-AWARE MULTIMODAL MEMORY WITH ADAPTIVE AGENTIC RETRIEVAL FOR MULTI-PARTY SPOKEN CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory enables agents to accumulate information and reason across sessions, yet existing research primarily focuses on dyadic text or image-text conversations, leaving long-term memory for multi-party spoken conversations underexplored. This setting requires preserving conversational content, identifying participants across sessions, and retaining who speaks to whom. To this end, we propose VoxPolyMem, an interaction-aware multimodal memory framework combining incremental speaker identification with a memory hierarchy comprising interaction memory, fact memory, and participant profiles. We formulate retrieval as sequential decision-making, where an agent rewrites queries and selects retrieval tools and memory layers based on accumulated evidence to address information gaps. We further introduce Evidence-Gain GRPO (EG-GRPO), which uses round-wise credit assignment to encourage complementary evidence acquisition. We also construct VoxPolyBench to evaluate memory evolution, personalized answering, memory retrieval and reasoning, and interaction reasoning and attribution in multi-party spoken conversations. VoxPolyMem achieves an overall score of 85.0 on VoxPolyBench, surpassing the strongest evaluated baseline by 23.6 points. On Mem-Gallery and H2HMem-Multi, it scores 89.6 and 74.4, respectively, exceeding the strongest evaluated public memory baselines by over 8 points each. These results highlight its potential for persistent, personalized assistance in multi-party multimodal interactions. Code and datasets are available at `https://voxpolymem.github.io/VoxPolyBench/demo/`.

1 INTRODUCTION

Long-term memory enables agents to retain information and reason across sessions (Hu et al., 2026c; Luo et al., 2026; Huang et al., 2026). Recent systems consolidate conversational facts, link related memories, and incorporate multimodal information (Chhikara et al., 2025; Xu et al., 2025; Liu et al., 2025; Feng et al., 2026). However, evaluations largely focus on dyadic text or image-text conversations (Wu et al., 2025; Bei et al., 2026). In-car, meeting, and household assistants often engage with multiple participants through speech, requiring them to remember information and interactions across sessions. Recent work extends memory evaluation to multi-party text or image-text histories (Yang et al., 2026) and explores multi-speaker speech understanding and voice-based personalization (Xiong et al., 2026; Al-Ratrout et al., 2026). Yet long-term memory for multi-party spoken conversations remains underexplored, with limited benchmarks for evaluating recurring speaker identification, participant-specific memories, and evolving interactions across sessions.

Constructing memory for these conversations requires more than preserving transcribed content (Inoue et al., 2025). It also requires extracting voice representations from audio and linking them to participant identities across sessions. Storing audio solely as ASR transcripts leaves these acoustic identity cues unmodeled. Moreover, existing memory methods organize conversational histories through temporal graphs and semantic abstraction (Rasmussen et al., 2025; Liu et al., 2026), but do not explicitly store interaction roles, such as who made a statement and to whom it was addressed. Preserving speaker identity and interaction roles supports the extraction of factual memories and

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.

participant profiles that capture individual preferences, commitments, and relationships. Retrieving supporting evidence poses a further challenge, as information may span sessions, memory layers, and modalities, with tools ranging from keyword search to cross-modal retrieval (Yeo et al., 2026). Fixed retrieval strategies may overlook information needs that emerge as evidence accumulates. Iterative retrieval enables multi-step evidence gathering (Trivedi et al., 2023), while learned search optimizes retrieval using terminal answer rewards (Jin et al., 2025). However, terminal rewards do not explicitly isolate the contribution of each retrieval round to acquiring new supporting evidence.

To address these challenges, we propose VoxPolyMem, an interaction-aware multimodal memory framework that combines incremental speaker identification with interaction memory, fact memory, and participant profiles. A directed interaction graph records who said what to whom. A dialogue window provides recent conversational context for extracting self-contained facts and participant attributes, while profiles link participant names and aliases to their voice representations. Building on this hierarchy, a trainable retrieval agent selects memory layers and tools and rewrites queries based on accumulated evidence and previous actions. We train the agent with Evidence-Gain GRPO (EG-GRPO), an adaptation of Group Relative Policy Optimization (GRPO) (Shao et al., 2024). At each retrieval state, EG-GRPO compares candidate actions through the coverage and ranking of newly acquired supporting evidence retained in the context. Previously retained evidence receives no additional credit, encouraging complementary retrieval across memory layers and modalities within a fixed budget.

To evaluate long-term memory in multi-party spoken conversations, we construct VoxPolyBench, comprising 1,527 QA pairs over 176 sessions and 18.9 hours of synthesized speech across 18 scenarios. Each scenario follows recurring participants across sessions, enabling evaluation of cross-session participant identification, memory evolution, personalized answering, retrieval and reasoning, and interaction attribution. Our contributions are threefold:

- We introduce VoxPolyBench, a benchmark for long-term memory in multi-party spoken conversations, covering recurring participant identification, memory evolution, personalized answering, retrieval and reasoning, and interaction attribution.
- We propose VoxPolyMem, integrating incremental speaker identification, interaction-aware hierarchical memory, and adaptive retrieval. Its retrieval agent selects memory layers and tools and rewrites queries, with EG-GRPO rewarding newly acquired supporting evidence across retrieval rounds.
- Experiments on VoxPolyBench, Mem-Gallery, and H2HMem-Multi demonstrate consistent improvements in overall answer quality. Ablations highlight the benefits of interaction-aware memory, while retrieval-policy comparisons show improved evidence recall with fewer retrieval rounds.

2 Related Work

2.1 Long-Term Memory Construction and Retrieval

Agent memory systems organize information from prior conversations and observations for subsequent retrieval and reasoning. Mem0 consolidates facts, A-Mem links contextual notes, and Zep maintains temporal knowledge graphs (Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025). LightMem and SimpleMem improve efficiency through staged consolidation and semantic compression (Fang et al., 2026; Liu et al., 2026). Multimodal systems introduce hierarchical, episodic, and semantic memories for image-text and audiovisual experiences (Liu et al., 2025; Feng et al., 2026; Lin et al., 2025; Long et al., 2025; Lian et al., 2026). AFA uses voice-based identification to separate users' memories (Al-Ratrout et al., 2026). For shared multi-party conversations, VoxPolyMem links recurring speaker identities to directed interaction relations, preserving who said what to whom when extracting facts and participant profiles.

Retrieval methods combine routing, reasoning, and policy learning. UniversalRAG routes queries across modalities and granularities, while IRCoT interleaves retrieval with reasoning (Yeo et al., 2026; Trivedi et al., 2023). Search-R1 and Memory-R1 use outcome rewards to optimize search or memory operations (Jin et al., 2025; Yan et al., 2025). LeTS combines process and outcome rewards, while Mem-T provides dense credit for memory operations (Zhang et al., 2025; Yue et al., 2026). InfoReasoner estimates retrieval gains through reductions in semantic uncertainty (Hu et al.,

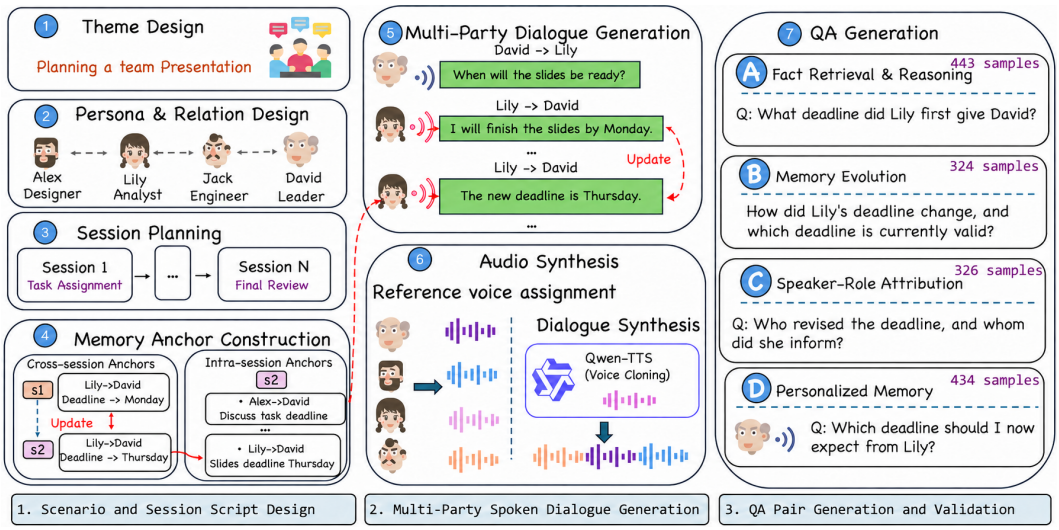


Figure 1: Construction of VoxPolyBench. Scenario plans and event anchors guide multi-party dialogue and speech generation, followed by QA construction and validation. Task counts follow the dataset taxonomy in Appendix A.2; result aggregation is specified in Appendix C.2.

2026b). EG-GRPO rewards the coverage and ranking of newly acquired supporting evidence retained after each action. Rewards are computed after context selection, with previously retained evidence excluded from both terms.

2.2 Long-Term Memory and Spoken Interaction Benchmarks

LoCoMo and LongMemEval evaluate long-term conversational recall and reasoning (Maharana et al., 2024; Wu et al., 2025). Mem-Gallery extends evaluation to image-text histories, while H2HMem, GroupMemBench, and EverMemBench cover multi-party memory settings (Bei et al., 2026; Zhu et al., 2026; Yang et al., 2026; Hu et al., 2026a). These settings do not jointly evaluate acoustic identity recovery and participant-dependent memory. For spoken interaction, ContextDialog tests conversational recall, MSU-Bench evaluates multi-speaker understanding, and Vox-Infinity studies long-context speech understanding (Kim et al., 2025; Wang et al., 2025; Anonymous, 2026). VoxPolyBench jointly evaluates recurring speaker identification, memory evolution, personalized answering, retrieval and reasoning, and speaker and addressee attribution over multi-session, multi-party spoken histories.

3 VoxPolyBench

VoxPolyBench comprises 18 scenarios across domains including telemarketing, meetings, in-car assistance, and household assistance. Each scenario contains a dialogue history spanning 8–12 sessions with recurring participants. In total, the benchmark contains 176 sessions, 18.9 hours of synthesized speech, and 1,527 QA pairs. The QA pairs cover memory evolution, personalized answering, retrieval and reasoning, and interaction attribution. Table 1 compares its scope with existing benchmarks, and Appendix A.2 provides detailed statistics.

We construct the benchmark in three stages, using event anchors to plan how information and participant interactions develop across sessions. Figure 1 summarizes the pipeline.

Stage 1: Hierarchical Scenario and Event Planning. We first define each scenario's theme and participant settings, including identities, roles, long-term attributes, and interpersonal relationships. We then plan session topics and event progression, specifying event anchors within and across sessions. Within-session anchors describe local events and participant interactions, while cross-session anchors track the persistence, updates, and conflicts of facts and preferences. Each anchor records the relevant participants, time, and event content; interaction anchors additionally specify speakers and addressees. Human reviewers check each session's anchors before dialogue generation and re-

vise any identified errors or implausible event transitions to maintain logical consistency and support natural dialogue generation.

Stage 2: Multi-Party Dialogue and Speech Generation. We use GPT-4.1 to generate multi-party dialogues from session topics and event anchors, conditioned on participant settings, prior event states, and subsequent anchors to maintain continuity across sessions. Speech synthesis uses a fixed reference voice for each participant across sessions, with telephone numbers and alphanumeric identifiers rendered for character-by-character pronunciation. We check synthesis duration and use ASR to identify transcription mismatches, identifier errors, and suspected repetitions. Targeted resynthesis, transcription rechecks, and manual spot checks address identified quality issues. Final evaluation with Whisper large-v3-turbo over all dialogue audio segments yields a corpus-level word error rate (WER) of 2.14% against normalized source texts. Appendix A.1 details the quality-control and evaluation procedures.

Stage 3: QA Generation and Validation. We construct QA pairs from event anchors and the generated dialogues, designing questions around event states and their changes, determining answers from dialogue content, and annotating the dialogue turns that support each answer. Questions examine both the integration and updating of memories across sessions and whom the information concerns, who expressed it, and to whom it was addressed. Human reviewers and an LLM check event consistency, participant and interaction assignments, question clarity, and evidence sufficiency. Correctable errors are revised and unsupported or ambiguous QA pairs removed, following Appendix A.1.

Table 1: Comparison of spoken-dialogue and memory benchmarks. ✓ denotes explicit evaluation, ○ partial or adjacent coverage, and – no targeted evaluation. Interaction attribution covers both speakers and addressees; personalization requires answers conditioned on the querying participant. Sources and comparison criteria are detailed in Appendix A.3.

Benchmark	**Modality**	Multi-party	Multi-session	Cross-session identity	Interaction attribution	Personalization	Memory update
ContextDialog	Speech	–	–	–	–	–	–
MSU-Bench	Speech	✓	–	–	○	–	–
AFA	Speech	–	○	✓	○	✓	○
Vox-Infinity	Speech	–	–	–	–	–	–
Mem-Gallery	Text+Image	–	✓	–	–	–	✓
H2HMem	Text+Image	✓	✓	✓	–	–	✓
VoxPolyBench	Speech	✓	✓	✓	✓	✓	✓

4 VoxPolyMem

Overview. As illustrated in Figure 2, VoxPolyMem connects input processing, memory construction, and query answering. For incoming multimodal conversations, online speaker identification assigns speaker identifiers to audio turns in Section 4.1. These identifiers and conversational content support the construction of interaction memory and the joint extraction of facts and participant profiles in Section 4.2. Given a query, the retrieval agent uses profile voice memory to identify spoken-query participants, then iteratively rewrites queries and selects memory layers and tools to gather supporting evidence for answering in Section 4.3. EG-GRPO trains this policy to acquire complementary evidence across retrieval rounds.

4.1 Online Speaker Identification

To distinguish recurring participants without a predefined speaker count, we maintain an online speaker memory. Acoustic matching assigns anonymous identifiers, and accumulated acoustic evidence reconciles fragmented identifiers. Names and aliases are associated later during LLM-based memory extraction in Section 4.2.

Speaker encoding. For audio x_i^{aud} at turn i, an ECAPA-TDNN encoder E_{spk} (Desplanques et al., 2020) produces the voiceprint embedding $\mathbf{e}_i$ in equation 1. The operator norm applies L_2 normalization, so inner products measure cosine similarity.

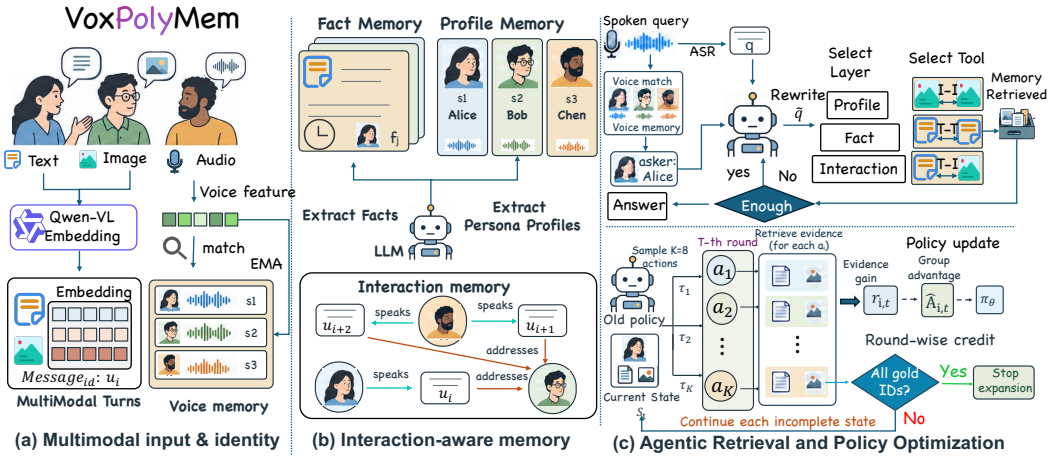


Figure 2: VoxPolyMem identifies recurring speakers, constructs interaction memory, facts, and participant profiles, and adaptively retrieves evidence across memory layers and modalities.

Incremental identity matching. We maintain a set $\mathcal{V}$ of entries $(k, \mathbf{v}_k)$, where k is an anonymous `speaker_id` and $\mathbf{v}_k$ its stored, normalized voiceprint embedding. The set is initially empty, so the first spoken turn creates an identifier with $\mathbf{e}_i$ as its voiceprint embedding. For each subsequent audio turn, we compare $\mathbf{e}_i$ with all stored embeddings using cosine similarity. Equation 1 selects the most similar entry, yielding identifier k^* and similarity score s_i. If $s_i \geq \tau_{\text{match}}$, we assign the turn to k^*. Otherwise, no stored speaker provides a sufficiently close match, so we create a new identifier initialized with $\mathbf{e}_i$. This procedure accommodates newly appearing speakers without requiring their number in advance. Non-audio turns leave $\mathcal{V}$ unchanged.

$$\mathbf{e}_i = \text{norm}\big(E_{\text{spk}}(x_i^{\text{aud}})\big)\,, \quad k^* = \underset{k:\,(k,\mathbf{v}_k)\in\mathcal{V}}{\arg\max}\ \mathbf{e}_i^\top \mathbf{v}_k, \qquad s_i = \mathbf{e}_i^\top \mathbf{v}_{k^*}. \tag{1}$$

Confidence-gated memory updates. To accommodate acoustic variation without reinforcing uncertain matches, we update a matched embedding by exponential moving average (EMA) only when $s_i \geq \tau_{\text{update}} > \tau_{\text{match}}$. The update in equation 2 uses weight α for the current observation. Scores between the thresholds associate the turn with an existing identifier but leave its embedding unchanged.

$$\mathbf{v}_{k^*} \leftarrow \text{norm}((1-\alpha)\mathbf{v}_{k^*} + \alpha \mathbf{e}_i)\,. \tag{2}$$

Acoustic variation may cause the same speaker to receive multiple provisional speaker IDs. After each session, we revisit these IDs using accumulated voiceprint evidence and conservatively merge those supported as belonging to the same speaker, while retaining their underlying voiceprint embeddings. This reduces identity fragmentation without requiring a participant roster or target speaker count. Appendix B.1 provides the merging criteria and configuration.

4.2 Hierarchical Interaction-Aware Multimodal Memory

We organize conversational information into interaction memory $\mathcal{M}_{\text{int}}$, fact memory $\mathcal{M}_{\text{fact}}$, and participant profiles $\mathcal{M}_{\text{prof}}$. Interaction memory preserves multimodal messages and their participant relations as source evidence. Fact memory captures self-contained statements grounded in these interactions, while participant profiles organize identity, background, and preferences around individual participants. Facts and profiles are jointly extracted from local conversational context.

Multimodal interaction memory. Each message retains its text, original multimodal attachments, timestamp, and session position. Text, image captions, and automatic speech recognition (ASR) transcripts from `whisper-turbo` support semantic extraction. Qwen3-VL-Embedding (Li et al., 2026) encodes text and images for content retrieval, separately from the voiceprint embeddings used for speaker identification. We represent interaction memory as a directed graph $\mathcal{M}_{\text{int}} = (\mathcal{P} \cup \mathcal{U}, \mathcal{E})$, where $\mathcal{P}$ and $\mathcal{U}$ contain participant and message nodes. The edge set records who speaks and to

whom:

$$\mathcal{E} = \bigcup_i \left(\left\{ \text{speaker}_i \xrightarrow{\text{speaks}} u_i \right\} \cup \left\{ u_i \xrightarrow{\text{addresses}} p \;\middle|\; p \in \mathcal{A}_i \right\} \right). \tag{3}$$

For message u_i, speaker_i denotes its speaker and $\mathcal{A}_i$ its addressee set. Addressees are inferred through the context-aware extraction described below. Unknown addressees introduce no participant-specific edge.

Context-aware fact and participant memory. An LLM processes the current turn and up to three preceding turns to associate anonymous `speaker_id`s with participant names and aliases while jointly extracting facts, participant attributes, addressees, and temporal references. Identifiers resolved to the same participant are linked through an identity mapping while retaining their voiceprint embeddings. The set $\mathcal{K}_p$ contains the identifiers associated with participant p, whose profile also records background and preferences. Fact memory stores self-contained statements with their speakers, addressees, and time. The records in equation 4 summarize these fields; image-related facts also retain their image identifiers. Each fact and profile retains `refer_id` links to its source dialogue turns.

$$\begin{aligned} \mathcal{M}_{\text{fact}} &= \left\{ \left(\text{content}_j,\, \text{speaker}_j,\, \text{addressee}_j,\, \text{time}_j \right) \right\}_j, \\ \mathcal{M}_{\text{prof}} &= \left\{ \left(\text{name}_p,\, \text{aliases}_p,\, \{(k, \mathbf{v}_k)\}_{k \in \mathcal{K}_p},\, \text{attributes}_p \right) \right\}_{p \in \mathcal{P}}. \end{aligned} \tag{4}$$

The index j ranges over extracted facts, whose fields record content, speaker, addressees, and time. Each participant profile stores a name, aliases, associated voiceprint embeddings, and attributes comprising background and preferences.

4.3 Trainable Agentic Memory Retrieval

Evidence-conditioned retrieval. Queries may require evidence across memory layers and modalities, with intermediate results guiding subsequent retrieval. For spoken queries, voiceprint matching against profiles identifies asker_q; unmatched speakers remain unknown. The state and action are

$$\begin{aligned} s_t &= (q, \text{asker}_q, C_{t-1}, H_{t-1}, b_t), \\ a_t &= (\ell_t, \rho_t, \widetilde{q}_t, f_t) \sim \pi_\theta(\cdot \mid s_t), \qquad \rho_t \in \mathcal{R}(\ell_t). \end{aligned} \tag{5}$$

Here, t indexes rounds. The state contains query q, asker identity, retained evidence C_{t-1}, action history H_{t-1}, and remaining budget b_t. Policy π_θ selects layer ℓ_t (interaction, fact, or profile), tool $\rho_t \in \mathcal{R}(\ell_t)$, rewritten query $\widetilde{q}_t$, and optional speaker/addressee filters f_t. Tools support text-to-text retrieval through vector similarity or BM25, text-to-image retrieval through image descriptions, and image-to-image retrieval through image embeddings. Reciprocal rank fusion (RRF), deduplication, and context selection update the retained evidence. An LLM identifies missing information to guide further retrieval, or answers when evidence is sufficient or three rounds are exhausted.

Evidence-Gain GRPO. To support adaptive retrieval, EG-GRPO adapts GRPO (Shao et al., 2024) to reward newly acquired supporting evidence retained in context. At each state, π_{old} samples $K = 8$ independent candidates sharing the pre-action context but not retrieved results. They form one GRPO group, indexed by i below.

After RRF fusion and deduplication, candidate i retains up to $k = 15$ memory records in C_i, including fact and profile contents. For reward calculation, interaction-turn IDs and `refer_id` links define the source-ID set U_i. Let $G_q \neq \emptyset$ contain annotated supporting IDs and B IDs represented in earlier retained contexts on this branch, initially empty. Then $D_i = (U_i \cap G_q) \setminus B$ contains newly retained supporting IDs. For $e \in U_i$, $p_i(e)$ is the earliest rank of a memory in C_i linked to e.

$$\begin{aligned} N_i &= \frac{1}{Z_q} \sum_{e \in D_i} \frac{1}{\log_2(p_i(e) + 1)}, \qquad Z_q = \sum_{j=1}^{\min(k, |G_q|)} \frac{1}{\log_2(j+1)}, \\ r_i &= \lambda \frac{|D_i|}{|G_q|} + (1 - \lambda) N_i, \qquad \widehat{A}_i = \frac{r_i - \mu}{\sigma + \delta}. \end{aligned} \tag{6}$$

In equation 6, coverage rewards newly retained supporting IDs, while N_i discounts them by memory rank using fixed scale Z_q. Each source ID contributes once, even if linked to multiple records. The

weight $\lambda \in [0, 1]$ balances coverage and ranking. The advantage $\widehat{A}_i$ uses the mean μ and standard deviation σ of this state's rewards, with $\delta > 0$ for stability.

Each candidate creates a successor $s_{t+1}^{(i)}$ with its selected evidence C_i, updated action history, and budget $b_{t+1}^{(i)} = b_t - 1$. During training, we expand

$$\mathcal{S}_{\text{next}}(s_t) = \left\{ s_{t+1}^{(i)} \;\middle|\; G_q \setminus U_i \neq \emptyset, \quad b_{t+1}^{(i)} > 0 \right\}. \tag{7}$$

Each successor in $\mathcal{S}_{\text{next}}$ independently samples another K actions. Unexpanded successors' actions remain in their current group. All groups update the shared policy through equation 8.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Big[\tfrac{1}{K} \sum_{i=1}^{K} \tfrac{1}{|a_i|} \sum_{n=1}^{|a_i|} \big\{ \min\big[w_{i,n} \widehat{A}_i,\ \text{clip}(w_{i,n}, 1-\epsilon, 1+\epsilon) \widehat{A}_i \big] - \beta \text{KL}_{i,n} \big\} \Big]. \tag{8}$$

Here, n indexes the $|a_i|$ action tokens, distinct from memory positions $p_i(e)$. The ratio $w_{i,n}$ compares current and old policies under the same state and action prefix; $\text{KL}_{i,n}$ measures divergence from a fixed reference policy. The coefficients ϵ and β control clipping and KL regularization. Advantages update only generated layer, tool, query, and filter tokens. Input, retrieved, and answer tokens are excluded; the answer model remains fixed. Gold annotations are training-only. Appendix B.3 details source ordering, history updates, and optimization.

5 Experiments

We evaluate VoxPolyMem on multi-party spoken and public image-text memory benchmarks, examining overall performance, component contributions, and retrieval policy optimization. Beyond synthesized speech in VoxPolyBench, the speaker tracker achieves attribution accuracies of 95.0% on IEMOCAP and 87.9% on AMI recordings with reference utterance boundaries (Appendix D.1).

5.1 Experimental Setup

Evaluation setup. We compare VoxPolyMem with ten text-based and multimodal baselines on the held-out test splits of VoxPolyBench, Mem-Gallery (Bei et al., 2026), and H2HMem-Multi (Zhu et al., 2026). On VoxPolyBench, all methods construct semantic memory from the same `whisper-turbo` ASR transcripts rather than the original dialogue scripts. VoxPolyMem and its ablations perform speaker identification automatically, while baselines receive ground-truth speaker identification labels. The public benchmarks retain their native text and visual inputs. All methods and ablations use GPT-4.1-mini as the answer model, with shared evaluation questions and rubric and a final-context budget of $k = 15$ retrieved records within each benchmark. GPT-4.1-mini and GPT-4.1 independently judge each answer. We average their scores per answer and then across test questions to obtain Score. Recall measures the fraction of annotated supporting dialogue turns represented in the final context, averaged over questions with nonempty annotations. It is reported on VoxPolyBench and Mem-Gallery, since H2HMem-Multi lacks these annotations. Both metrics are multiplied by 100 and rounded to one decimal. Appendix C details baselines, metrics, and evaluation protocols, while Appendix D.2 reports the human audit.

RL training. We train an 8B Qwen-VL retrieval policy for one epoch with EG-GRPO. GPT-4.1-mini handles memory extraction. Training uses approximately 900 QA instances: 300 from Mem-Gallery, 400 from LoCoMo (Maharana et al., 2024), and 200 from VoxPolyBench. These datasets annotate supporting dialogue turns for each question, providing supervision for EG-GRPO. Table 5 reports evaluation split counts. The annotations determine rewards and training-time stopping based on current evidence coverage and are unavailable at inference.

5.2 Main Results

To assess multi-party spoken and image-text memory, we evaluate VoxPolyMem on VoxPolyBench and the public Mem-Gallery and H2HMem-Multi benchmarks. VoxPolyMem leads all three, with a weighted average of 86.6 versus 68.1 for the strongest external baseline (Table 2).

Table 2: Main results across three memory benchmarks (↑). For VoxPolyBench, RR/EC/P/IA denote the retrieval and reasoning, evolution and conflict, persona, and interaction attribution subsets, respectively. For Mem-Gallery, E&A/R/KM denote the extraction and adaptation, reasoning, and knowledge management subsets, respectively. For H2HMem-Multi, A/R/Rec. denote the memory application, memory reasoning, and memory recall subsets, respectively. Aver denotes the mean score within each benchmark. Average weights benchmark scores by their test-question counts. Bold and underlined values mark the best and second-best results.

	VoxPolyBench					**Mem-Gallery**				**H2HMem-Multi**				**Average**
Method	RR	EC	P	IA	**Aver**	E&A	R	KM	**Aver**	A	R	Rec.	**Aver**	**Score**
mem0 (2025)	79.1	79.6	40.5	42.6	60.4	50.6	61.2	81.9	60.6	71.3	29.8	53.9	54.6	60.2
A-Mem (2025)	76.1	77.3	34.8	32.4	55.3	42.7	55.5	72.5	53.0	70.0	35.7	48.5	54.1	54.1
LightMem (2026)	83.5	80.2	37.0	38.1	59.9	40.2	41.8	79.0	49.4	67.7	32.7	59.3	55.4	54.5
AUGUSTUS (2025)	77.1	78.9	38.1	32.2	56.8	77.1	68.0	74.0	73.8	77.3	48.2	66.5	66.0	65.6
UniversalRAG (2026)	77.8	80.1	39.6	35.1	58.3	77.1	80.7	75.7	77.8	62.0	**58.3**	72.5	64.0	68.1
M2A (2026)	82.7	80.2	37.6	45.0	61.3	49.3	38.2	77.2	52.4	56.0	52.6	**77.0**	61.1	57.0
MemVerse (2025)	84.7	87.3	37.7	32.3	60.7	71.2	63.2	75.7	69.9	67.0	39.4	64.9	58.7	65.0
NGMemory (2025)	62.9	66.5	40.2	44.3	53.3	75.5	62.4	78.8	72.5	81.3	24.7	35.3	52.0	62.5
HippoMM (2025)	62.7	47.5	34.5	61.3	51.2	46.4	67.1	79.0	59.6	63.3	21.9	42.2	45.6	54.9
M3-Agent (2025)	70.5	68.4	25.2	40.0	50.7	68.8	55.8	81.1	67.9	49.7	3.1	18.1	27.4	57.6
w/o RL	85.5	90.8	75.9	86.0	84.0	87.9	83.9	85.9	86.3	85.0	52.0	72.1	72.1	84.4
VoxPolyMem	**86.4**	**92.1**	**76.1**	**87.5**	**85.0**	**91.4**	**87.3**	**88.5**	**89.6**	**88.1**	54.1	73.7	**74.4**	**86.6**

Table 3: Component ablations across three benchmarks. Ablated variants use w/o RL as the base model, with VoxPolyMem (full) included as a reference. Bold and underlined values mark the best and second-best results.

	VoxPolyBench		**Mem-Gallery**		**H2HMem-Multi**
Variant	Score	Recall	Score	Recall	Score
VoxPolyMem (full)	**85.0**	**87.4**	**89.6**	**93.6**	**74.4**
w/o RL	84.0	85.2	86.3	88.1	72.1
w/o hierarchical memory	79.8	81.4	84.9	88.9	63.6
w/o interaction relations	78.6	81.2	85.1	88.0	67.3
Single-round retrieval	76.1	74.8	85.5	87.7	60.0
w/o query rewriting	83.2	85.0	86.0	88.5	70.8

On VoxPolyBench, VoxPolyMem achieves an overall score of 85.0, exceeding the strongest external baseline by 23.6 points, and leads all four task subsets. Personalized answering and interaction attribution show particularly strong results, scoring 76.1 and 87.5, respectively. These gains are consistent with interaction-aware memory that preserves speakers, addressees, and participant attributes, allowing retrieval to filter memories by participant roles and interaction relations. Its advantages also extend to retrieval and reasoning and memory evolution, suggesting that interaction context supports broader memory use beyond explicit personalization and attribution questions.

On the public Mem-Gallery and H2HMem-Multi benchmarks, VoxPolyMem achieves overall scores of 89.6 and 74.4, exceeding the strongest external baselines by 11.8 and 8.4 points, respectively. It leads all three Mem-Gallery task subsets and memory application on H2HMem-Multi, supporting the applicability of hierarchical memory and adaptive retrieval to general multimodal memory tasks. Notably, the w/o RL variant already exceeds all external baselines on each benchmark, achieving a weighted average of 84.4. This indicates that the memory structure and agentic retrieval process are effective without additional RL, while policy optimization provides further overall improvements.

5.3 Component Ablations

Table 3 evaluates memory construction and retrieval components using w/o RL as the base model, with full VoxPolyMem as a reference.

Removing fact memory and participant profiles restricts retrieval to original interaction records and lowers scores across all benchmarks, most on H2HMem-Multi, from 72.1 to 63.6. Removing interaction edges and relation-based filtering while retaining speaker identities affects VoxPolyBench most, reducing its score from 84.0 to 78.6. These patterns suggest that memory abstraction supports access to dispersed information, while explicit interaction structure helps distinguish evidence associated with different participants.

Single-round retrieval causes larger declines than retaining multiple rounds with the original query, indicating greater sensitivity to iterative retrieval than to query rewriting without RL. Follow-up searches allow the agent to gather evidence missing from earlier results, even when the query remains unchanged. RL improves scores from 86.3 to 89.6 on Mem-Gallery and from 72.1 to 74.4 on H2HMem-Multi. These gains support optimizing the joint retrieval policy to coordinate memory-layer selection, tool use, and query rewriting based on accumulated evidence.

5.4 Comparison of Retrieval Policy Optimization Methods

Table 4 compares reward design and credit assignment under fixed memory, retrieval tools, and inference budget. We adapt Search-R1 (Jin et al., 2025) and MoT-GRPO from Mem-T (Yue et al., 2026) to our retrieval framework. Search-R1 uses a terminal answer reward, while MoT-GRPO combines tree-structured rollouts with evidence coverage and answer rewards. Terminal-coverage GRPO rewards final cumulative evidence coverage with a shared trajectory-level advantage. EG-GRPO instead assigns state-specific rewards and advantages for the coverage and ranking of newly acquired supporting evidence retained after each action.

Table 4: Comparison of retrieval policy optimization methods. Rounds denotes the mean number of executed retrieval rounds per question, excluding answer generation. Bold and underlined values mark the best and second-best results.

	VoxPolyBench			Mem-Gallery			H2HMem-Multi	
Method	Score ↑	Recall ↑	Rounds ↓	Score ↑	Recall ↑	Rounds ↓	Score ↑	Rounds ↓
w/o RL	84.0	85.2	1.6	86.3	88.1	1.7	<u>72.1</u>	1.4
Search-R1	81.3	82.3	1.6	86.1	90.1	1.6	71.9	<u>1.3</u>
Terminal-coverage GRPO	83.1	<u>86.7</u>	<u>1.4</u>	86.4	89.7	<u>1.5</u>	71.6	1.5
MoT-GRPO	<u>84.1</u>	86.6	<u>1.4</u>	<u>87.1</u>	<u>91.5</u>	1.8	72.0	1.6
EG-GRPO (ours)	**85.0**	**87.4**	**1.2**	**89.6**	**93.6**	**1.3**	**74.4**	**1.2**

Compared with w/o RL, Search-R1 lowers answer scores across all three benchmarks. Terminal-coverage GRPO improves recall on both annotated benchmarks without consistent answer improvements, showing that higher evidence coverage does not necessarily yield better answers. MoT-GRPO improves both metrics on VoxPolyBench and Mem-Gallery. EG-GRPO leads all three benchmarks in answer scores and both annotated benchmarks in recall. On Mem-Gallery, it raises answer scores from 87.1 to 89.6 and recall from 91.5 to 93.6 compared with MoT-GRPO.

These improvements accompany fewer retrieval rounds. EG-GRPO uses an average of 1.2–1.3 rounds across benchmarks, compared with 1.4–1.7 for w/o RL and 1.4–1.8 for MoT-GRPO. Its coverage reward encourages complementary searches, while its ranking reward favors newly acquired supporting evidence near the top of the selected context. These results support the overall training design for improving evidence acquisition and answering with fewer retrieval rounds.

6 Conclusion

To support persistent memory of participant identities and interactions across spoken conversations, we propose VoxPolyMem, combining incremental speaker identification, interaction-aware hierarchical memory, and adaptive retrieval. EG-GRPO trains the retrieval policy using round-wise evidence gains. We also introduce VoxPolyBench to evaluate memory over multi-session spoken interactions. Experiments on VoxPolyBench and two public image-text benchmarks support the effectiveness of the memory structure and agentic retrieval, with further improvements from policy optimization. Future work will extend evaluation to real-world spoken interactions with greater acoustic and conversational diversity. Further limitations appear in Appendix E.

AI USE STATEMENT

LLMs were used for benchmark construction and the memory and evaluation pipelines, as described in Sections 3–5. Generative AI tools also assisted code development. The authors take responsibility for the final manuscript and research artifacts.

ETHICS STATEMENT

VoxPolyBench uses generated dialogues and synthesized speech. This controlled setting does not establish robustness across real speakers, accents, or recording conditions. Voiceprint embeddings, identity links, and participant profiles can contain sensitive information and enable unwanted tracking or disclosure. Real-world deployment should require informed consent, restricted access, limited retention, and mechanisms to correct or delete stored information. Benchmark performance should not be interpreted as authorization for surveillance or high-stakes identity decisions.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe benchmark construction and the method. Appendices A–D provide data validation, implementation settings, evaluation protocols, and additional experiments. Reproducibility with hosted models also depends on API versions and sampling variability.

REFERENCES

Mohammad Al-Ratrout, Pavan Uttej Ravva, Shayla Sharmin, Aditya Raikwar, Ju Young Shin, and Roghayeh Leila Barmaki. AFA: Identity-aware memory for preventing persona confusion in multi-user dialogue. *arXiv preprint arXiv:2604.25022*, 2026. URL https://arxiv.org/abs/2604.25022.

Anonymous. Vox-Infinity: Benchmarking the limits of long-context spoken language models. Unpublished manuscript, 2026.

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, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

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

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650. URL https://doi.org/10.21437/Interspeech.2020-2650.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.18866.

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, 2025. URL https://doi.org/10.5281/zenodo.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, 2026a. URL https://arxiv.org/abs/2602.01313.

Senkang Hu, Yong Dai, Yuzhi Zhao, Yihang Tao, Yu Guo, Zhengru Fang, Sam Tak Wu Kwong, and Yuguang Fang. Optimizing agentic reasoning with retrieval via synthetic semantic information gain reward, 2026b. URL https://arxiv.org/abs/2602.00845.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the age of ai agents, 2026c. URL https://arxiv.org/abs/2512.13564.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, Shanglin Wu, Ruiyao Xu, Liangwei Yang, Rui Yang, Wooseong Yang, Chin-Yuan Yeh, Hanrong Zhang, Haozhen Zhang, Siqi Zhu, Henry Peng Zou, Wanjia Zhao, Song Wang, Wujiang Xu, Zixuan Ke, Zheng Hui, Dawei Li, Yaozu Wu, Langzhou He, Chen Wang, Xiongxiao Xu, Baixiang Huang, Juntao Tan, Shelby Heinecke, Huan Wang, Caiming Xiong, Ahmed A. Metwally, Jun Yan, Chen-Yu Lee, Hanqing Zeng, Yinglong Xia, Xiaokai Wei, Ali Payani, Yu Wang, Haitong Ma, Wenya Wang, Chenguang Wang, Yu Zhang, Xin Eric Wang, Yongfeng Zhang, Jiaxuan You, Hanghang Tong, Xiao Luo, Xue Liu, Yizhou Sun, Wei Wang, Julian McAuley, James Zou, Jiawei Han, Philip S. Yu, and Kai Shu. A survey of agent memory in the second half: Towards self-evolving and long-horizon agents, 2026. URL https://arxiv.org/abs/2602.06052.

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. An LLM benchmark for addressee recognition in multi-modal multi-party dialogue, 2025. URL https://arxiv.org/abs/2501.16643.

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516.

Heeseung Kim, Che Hyun Lee, Sangkwon Park, Jiheum Yeom, Nohil Park, Sangwon Yu, and Sungroh Yoon. Does your voice assistant remember? analyzing conversational context recall and utilization in voice interaction models. *arXiv preprint arXiv:2502.19759*, 2025. URL https://arxiv.org/abs/2502.19759.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. *arXiv preprint arXiv:2603.01455*, 2026. URL https://arxiv.org/abs/2603.01455. Accepted at ACL 2026.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Li, and Yiran Chen. HippoMM: Hippocampal-inspired multimodal memory for long audiovisual event understanding, 2025. URL https://arxiv.org/abs/2504.10739. Accepted at CVPR 2026 Findings.

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

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, Yi Yu, Shuyue Hu, Botian Shi, and Ding Wang. MemVerse: Multimodal memory for lifelong learning agents, 2025. URL https://arxiv.org/abs/2512.03627.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory, 2025. URL https://arxiv.org/abs/2508.09736.

Jinghao Luo, Yuchen Tian, Chuxue Cao, Ziyang Luo, Hongzhan Lin, Kaixin Li, Chuyi Kong, Ruichao Yang, and Jing Ma. From storage to experience: A survey on the evolution of llm agent memory mechanisms, 2026. URL https://arxiv.org/abs/2605.06716.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024. URL https://arxiv.org/abs/2402.17753.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. MSU-Bench: Towards understanding the conversational multi-talker scenarios. *arXiv preprint arXiv:2508.08155*, 2025. URL https://arxiv.org/abs/2508.08155.

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://arxiv.org/abs/2410.10813.

Chenxu Xiong, Dongming Shen, Yuzhi Tang, Wentao Ma, Mu Li, and Alex Smola. MSI-Bench: Evaluating multi-speaker voice interaction for collaborative AI agents, 2026. URL https://arxiv.org/abs/2609.24812.

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 38*, pp. 20004–20031, 2025. doi: 10.52202/085713-0593. URL https://doi.org/10.52202/085713-0593.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025. URL https://arxiv.org/abs/2508.19828.

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.

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

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. Mem-T: Densifying rewards for long-horizon memory agents. *arXiv preprint arXiv:2601.23014*, 2026. URL https://arxiv.org/abs/2601.23014v2.

Qi Zhang, Shouqing Yang, Lirong Gao, Hao Chen, Xiaomeng Hu, Jinglei Chen, Jiexiang Wang, Sheng Guo, Bo Zheng, Haobo Wang, and Junbo Zhao. LeTS: Learning to think-and-search via process-and-outcome reward hybridization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5109–5122. Association for Computational Linguistics, November 2025. doi: 10.18653/v1/2025.emnlp-main.257. URL https://aclanthology.org/2025.emnlp-main.257/.

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.

A Benchmark Details

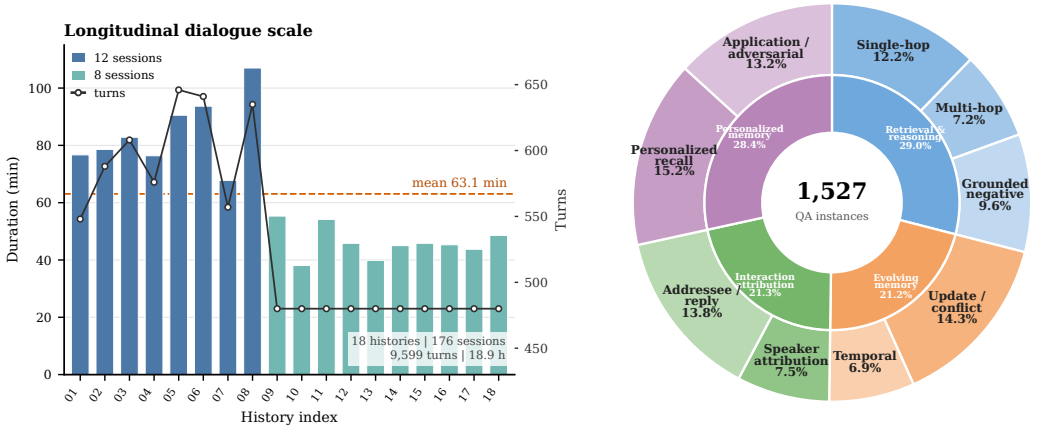
A.1 Construction and Quality Control

Quality control covers session anchors and the resulting QA annotations. Before dialogue generation, human reviewers inspect each session's anchors for consistency with participant settings and the planned event progression. After generation, QA pairs undergo consistency checks by both human reviewers and an LLM.

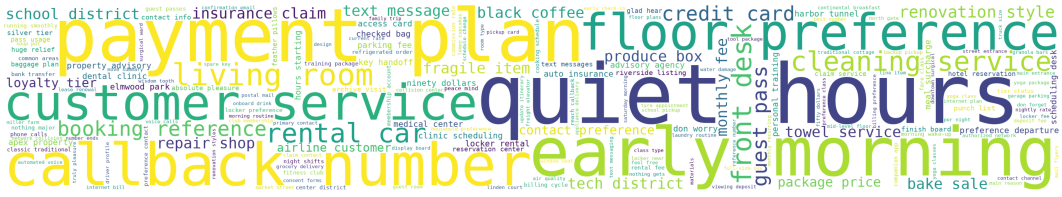
The checks address four requirements. Realized events must agree with the anchors and their temporal updates. Speaker and addressee assignments must agree with the dialogue. Questions must be clear, and reference answers must be supported by the conversation. Annotated evidence must be sufficient to support the answer. Correctable errors are revised; QA pairs with unresolved ambiguity or insufficient support are removed. These checks concern dataset construction, separately from the human audit of model-answer scores in Appendix D.2.

Speech synthesis quality control. We assign each participant a fixed reference voice across sessions and convert telephone numbers and alphanumeric identifiers to character-by-character pronunciations. Synthesis uses short reference clips and a target duration estimated from text length. We generate up to three candidates, preferring one within the duration tolerance and otherwise retaining the candidate with the smallest duration deviation. Duration serves as an auxiliary screen for potential abnormal lengthening or truncation. Post-synthesis ASR checks flag normalized text similarity below 0.85, telephone-number or identifier mismatches, and suspected repetitions. During construction, selected problematic samples undergo targeted resynthesis and transcription rechecks, supplemented by manual spot listening.

Speech transcription consistency. We independently transcribe all 9,599 final dialogue audio segments across the 18 scenarios using Whisper large-v3-turbo, with English recognition, temperature 0, and beam size 5. No reference texts or participant-name hints are supplied to the recognizer. Each reference is the complete source text used to synthesize the corresponding audio segment. We apply identical normalization to references and ASR outputs: curly apostrophes are mapped to straight apostrophes, em dashes to spaces, and non-breaking hyphens to standard hyphens, followed by Whisper's EnglishTextNormalizer to normalize case, punctuation, abbreviations, and number expressions. We do not correct transcriptions using the references.



(a) Longitudinal dialogue scale. **(b)** Hierarchical memory-task taxonomy.



(c) Dialogue-domain coverage across 9,599 turns.

Figure 3: VoxPolyBench dataset statistics. (a) Duration and turn count for 18 case-level histories, comprising 176 sessions and 18.9 hours of dialogue audio; cases are identified only by indices 01–18. (b) Distribution of the 1,527 QA instances in VoxPolyBench: the inner ring shows four coarse task families and the outer ring shows nine fine-grained types. (c) Case-balanced content terms extracted exclusively from dialogue turns after removing participant names, identifiers, quantities, and conversational filler.

We compute word-level edit distances for individual segments and aggregate substitution, deletion, and insertion counts over the corpus:

$$\mathrm{WER} = \frac{\sum_{i=1}^{M}(S_i + D_i + I_i)}{\sum_{i=1}^{M} N_i}, \tag{9}$$

where M is the number of evaluated segments, S_i, D_i, and I_i are the respective edit counts, and N_i is the number of normalized reference words in segment i. The resulting corpus-level WER is 2.14%; it is weighted by reference word count rather than averaged equally across segments. This metric measures agreement between automatic transcriptions and synthesis source texts, rather than a human-verified synthesis error rate.

A.2 Dataset Statistics

Longitudinal dialogue structure. VoxPolyBench contains 18 case-level histories spanning 176 sessions, 9,599 dialogue turns, and 18.9 hours of synthesized speech. Each case follows a recurring group of participants over 8–12 sessions and averages 63.1 minutes of dialogue, while the average session lasts 6.5 minutes. The resulting evaluation unit is a longitudinal multi-party history rather than an isolated recording or a collection of independent utterances. Figure 3(a) reports the case-level duration and turn distribution, and Figure 3(c) summarizes domain coverage using case-balanced dialogue terms.

Memory-task composition. VoxPolyBench contains 1,527 QA instances organized under a hierarchical taxonomy. Retrieval and reasoning accounts for 443 instances (29.0%), evolving memory for 324 (21.2%), interaction attribution for 326 (21.3%), and personalized memory for 434 (28.4%). The nine fine-grained task types cover single- and multi-hop retrieval, grounded negatives, memory

updates and conflicts, temporal reasoning, source-speaker attribution, addressee attribution, personalized recall, and application or adversarial questions. This distribution preserves conventional retrieval and temporal tasks while assigning substantial coverage to participant-dependent memory, as visualized in Figure 3(b).

A.3 Comparison with Existing Benchmarks

Table 1 separates three properties that are often conflated: spoken-dialogue understanding, multi-user personalization, and long-term multi-party memory. Existing speech benchmarks primarily evaluate within-session understanding or long-context recall without persistent participant identity. AFA routes independently interacting users to isolated persona memories, but does not represent a shared multi-party dialogue or evaluate structured long-horizon memory reasoning. Conversely, multimodal memory benchmarks provide longitudinal histories and richer memory operations, but generally expose participant labels and do not jointly evaluate spoken identity, source-speaker attribution, addressee attribution, and query-participant conditioning. VoxPolyBench combines these properties with memory evolution, cross-session evidence integration, and gold evidence provenance. To our knowledge, it is the first benchmark for long-term memory reasoning over multi-session, multi-party spoken conversations.

The comparison draws on ContextDialog (Kim et al., 2025), MSU-Bench (Wang et al., 2025), AFA (Al-Ratrout et al., 2026), the unpublished Vox-Infinity manuscript (Anonymous, 2026), MemGallery (Bei et al., 2026), and H2HMem (Zhu et al., 2026). Interaction attribution receives full coverage only when both source-speaker and addressee attribution are evaluated; coverage of only one is marked partial. The table summarizes targeted capabilities rather than ranking overall benchmark quality. Gold supporting evidence and cross-session evidence integration remain relevant evaluation properties, discussed here rather than shown as separate columns.

B Implementation Details

B.1 Acoustic Speaker Tracking

Online configuration. The matching and EMA-update thresholds are $\tau_{\text{match}} = 0.35$ and $\tau_{\text{update}} = 0.40$, with update weight $\alpha = 0.05$. Per-turn voiceprint embeddings and assignments are retained for subsequent identity reconciliation.

Candidate comparison. After each session and at the end of the input stream, utterances observed so far are grouped by their original speaker identifiers. Candidate pairs are scored using cross-group cosine similarities. Only a group at most one-third the size of its target is eligible to merge. A multi-utterance group's score is the median cross-group similarity. A singleton's score is its mean similarity to the five closest utterances in the target group.

Merging criteria. Multi-utterance groups must be mutual best matches, with sufficient separation from alternative candidates. Their cross-group similarity is assessed relative to within-group cohesion, and they must share at least half of the sessions represented by the group with fewer sessions. Singletons require a target containing at least ten utterances, a similarity score of at least 0.35, and sufficient separation from the runner-up when one exists.

The highest-scoring eligible pair is merged, candidates are recomputed, and the process stops when no eligible pair remains. The resulting alias mapping preserves original voiceprint embeddings and online EMA entries.

B.2 Memory and Retrieval Configuration

Acoustic reconciliation and context-aware name association produce identity mappings while preserving the underlying voiceprint embeddings. The former links fragmented speaker identifiers, while the latter associates them with participant names and profiles. Filters are specified through action parameters or resolved from the asker identity. Empty filters impose no participant constraints, and automatically supplied identity conditions are excluded from the policy loss. The final-context limit of $k = 15$ applies after RRF fusion, deduplication, and context selection, rather than separately to each tool call.

B.3 Evidence-Gain GRPO Details

Source evidence and ranking. The budget $k = 15$ limits retained memory records after RRF fusion and deduplication. These records, including extracted facts and profiles, form the answer context C_i. Source-ID mapping is used for reward calculation and does not replace these memories with source-turn text. We obtain U_i from retained interaction-turn IDs and fact/profile `refer_id` links, without an additional source-ID truncation. Multiple source IDs linked to one memory share its rank. If an ID occurs in several memories, its earliest memory rank defines $p_i(e)$, and it contributes only once. Sibling candidates are evaluated separately.

The denominators $|G_q|$ and Z_q provide fixed scales for a question. Because several supporting IDs may share one memory rank, N_i is a rank-weighted evidence gain rather than standard NDCG, and is not necessarily bounded by one. Candidates share the pre-action history and denominators, and their advantages are standardized within the current state. Empty retained contexts receive zero reward. Training requires nonempty supporting-turn annotations.

Branch history and expansion. The successor history is $B_i^{\text{next}} = B \cup U_i$, recording source IDs represented in retained memories. A source ID returned only through discarded memories is not added to this history and remains eligible for reward when subsequently retained. Siblings share the pre-action history but never contribute to each other's histories.

Each successor reduces the round budget by one. Training expansion stops when the current retained memories cover all supporting IDs, $G_q \subseteq U_i$, or the budget is exhausted. Historical coverage alone does not stop a branch. The completing action remains in its current training group. At inference, stopping uses evidence sufficiency and the round budget without gold annotations.

Policy optimization. The state-specific advantages in equation 6 enter the objective in equation 8. Let $a_{i,n}$ be token n of action i, and $a_{i,<n}$ its preceding tokens. The probability ratio is $w_{i,n} = \pi_\theta(a_{i,n} \mid s_t, a_{i,<n})/\pi_{\text{old}}(a_{i,n} \mid s_t, a_{i,<n})$, and $\text{KL}_{i,n} = D_{\text{KL}}[\pi_\theta(\cdot \mid s_t, a_{i,<n}) \| \pi_{\text{ref}}(\cdot \mid s_t, a_{i,<n})]$. The reference policy π_{ref} is fixed. Automatically supplied filters and all input or retrieved tokens are excluded from the action loss. The expectation in equation 8 covers sampled states and candidate groups. The evidence reward supervises retrieval actions, while the answer model remains fixed.

Example. Suppose $G_q = \{g_1, g_2\}$ and $B = \{g_1\}$. If three retained memory records support g_1, neither gold ID, and g_2, respectively, then $D_i = \{g_2\}$ and $p_i(g_2) = 3$. Coverage is $1/2$, and $N_i = (1/\log_2 4)/(1 + 1/\log_2 3) \approx 0.307$. If the memory supporting g_2 is discarded, g_2 is absent from U_i and remains eligible for reward in a later round.

C Evaluation Protocols

C.1 Data Splits and Baselines

Table 5 provides category-level QA counts for the three evaluation benchmarks. The training sources and shared evaluation settings are described in Section 5.1.

Baselines. The text-based memory baselines are mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025), and LightMem (Fang et al., 2026). The multimodal retrieval and memory baselines are AUGUSTUS (Jain et al., 2025), UniversalRAG (Yeo et al., 2026), M2A (Feng et al., 2026), MemVerse (Liu et al., 2025), NGMemory (Fisher, 2025), HippoMM (Lin et al., 2025), and M3-Agent (Long et al., 2025).

C.2 Metrics and Aggregation

Score aggregation. Within each benchmark, category scores are weighted by their test-question counts. The cross-benchmark Average uses the total test counts in Table 5 as weights. Aggregates, differences, rankings, and relative reductions are computed before display rounding. A dash denotes an unavailable measurement.

Table 5: Training and test QA counts. Total includes both splits; H2HMem-Multi is used only for testing. Category abbreviations follow Table 2; test counts determine the weights for score aggregation.

Benchmark	Category	Total	Train	Test
VoxPolyBench	RR	443	60	383
	EC	324	40	284
	P	434	60	374
	IA	326	40	286
	Total	**1,527**	**200**	**1,327**
Mem-Gallery	E&A	862	167	695
	R	503	103	400
	KM	346	30	316
	Total	**1,711**	**300**	**1,411**
H2HMem-Multi	A	75	0	75
	R	49	0	49
	Rec	51	0	51
	Total	**175**	**0**	**175**

Task grouping. Temporal reasoning belongs to EC in VoxPolyBench, following Figure 3. Mem-Gallery's nine tasks are grouped into three dimensions and aggregated by QA count. E&A comprises Factual Retrieval, Visual-centric Search, and Test-Time Learning. R comprises Temporal Reasoning, Visual-centric Reasoning, and Multi-entity Reasoning. KM comprises Knowledge Resolution, Conflict Detection, and Answer Refusal. The dimension names and abbreviations follow Table 2.

Evidence recall. Recall is measured after final-context deduplication and truncation. Retained facts and profiles contribute the source turn IDs linked through their `refer_id` fields, with each source turn ID counted once. Gold supporting-turn annotations are held fixed across variants. The per-question definition and averaging rule are given in Section 5.1.

C.3 LLM-AS-A-JUDGE PROTOCOL

The judges specified in Section 5.1 independently receive the same prompt below. Each request consists of a single user message containing the question, reference answer, and system answer.

Judge 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 have), 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>"}
```

D Additional Experiments

D.1 Speaker Identity Evaluation across Three Datasets

Protocol. We evaluate anonymous speaker-ID tracking on the 18 VoxPolyBench cases (9,599 turns), IEMOCAP Sessions 4–5 (4,273 utterances), and five AMI evaluation scenario chains (1,311 non-overlapping utterances). All variants reuse frozen ECAPA embeddings and oracle utterance boundaries, so the study isolates identity tracking rather than end-to-end diarization. Reference speaker labels and speaker counts are used only for scoring; inference receives neither a participant roster nor a target number of speakers.

Component analysis. Table 6 compares the static tracker with cumulative additions of confidence-gated EMA updates and identity reconciliation across all three datasets. EMA updates improve speaker attribution accuracy on all datasets, while identity reconciliation provides further gains, most notably on IEMOCAP, from 85.0% to 95.0%.

Table 6: Cumulative ablation of the shared anonymous speaker tracker. Each cell reports speaker attribution accuracy (%) after optimal permutation matching, with macro ARI $\times 100$ in parentheses. Neither reference identities nor the number of speakers is available during inference.

Variant	VoxPolyBench ↑	IEMOCAP ↑	AMI ↑
Static prototype	90.3 (90.0)	57.4 (35.7)	70.3 (58.6)
+ online EMA	97.5 (98.1)	85.0 (71.4)	87.3 (81.4)
+ identity reconciliation	**97.7 (98.4)**	**95.0 (80.8)**	**87.9 (82.4)**

D.2 Human Audit of LLM-Judge Scores

We compare aggregate LLM-judge and human scores on VoxPolyBench and Mem-Gallery. We sample 100 responses from each benchmark's complete w/o RL evaluation output using the fixed seed 20260920. Sampling preserves the judge-score distribution in proportion to the full result pool and distributes the selected items across the native task categories. Table 7 reports the mean LLM-judge score, the mean human score, and their signed and absolute gaps. All scores are measured on the same five-level rubric and multiplied by 100.

Table 7: Human audit of LLM-judge scores. Δ is Human $-$ LLM, and $|\Delta|$ is the absolute mean-score gap.

Benchmark	N	LLM	Human	Δ	$\lvert\Delta\rvert$
VoxPolyBench	100	83.3	85.0	+1.8	1.8
Mem-Gallery	100	86.0	87.3	+1.3	1.3
Overall	200	84.6	86.1	+1.5	1.5

The sampled LLM means differ from the w/o RL test scores of 84.0 on VoxPolyBench and 86.3 on Mem-Gallery by 0.8 and 0.3 points, respectively. Human means are slightly higher on both benchmarks, while the absolute gaps remain within 1.8 points. Across the 200 sampled responses, the aggregate LLM and human means are 84.6 and 86.1, respectively. Similar means do not establish item-level agreement between human and automatic scores.

E Limitations

VoxPolyBench uses generated dialogues and synthesized speech, with human review of event plans and QA pairs. Constructing challenging multi-hop questions from existing long conversations requires reliable links between dispersed evidence and unambiguous answers, which are difficult to ensure through automatic generation alone. Our current construction therefore uses planned events and human validation to control question and evidence quality. Future work will explore question generation and validation over transcripts of real long conversations, paired with their original audio, to broaden coverage of complex reasoning and real-world spoken interactions.