CSMEM: COGNITIVE STATE MEMORY FOR MULTIMODAL LONG-TERM DIALOGUE ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal long-term dialogue memory asks assistants to answer later questions from past image-text interactions. Existing multimodal memory methods preserve useful content and evidence through textualized records, raw multimodal memories, or graph-structured representations. However, these methods primarily preserve what appeared, rather than explicitly modeling the interaction state that gives those records meaning in context, including how references are resolved, alternatives are compared, decisions evolve, and claims are supported. To address this limitation, we introduce *Cognitive State Memory* (CSMem), a training-free memory framework. Inspired by situation-model accounts of comprehension, CSMem represents the structured state established within each coherent interaction rather than storing only its surface image-text records. It connects this state with retrieval cues and raw multimodal evidence to locate relevant interactions and generate grounded answers. Experiments on Mem-Gallery and LoCoMo show that CSMem outperforms textualized and raw multimodal methods, achieving up to an 11.3% relative improvement in exact match. Our code will be made publicly available soon.

1 INTRODUCTION

Multimodal long-term dialogue memory is the task of maintaining and querying a user's historical image-text interactions across sessions (Bei et al., 2026; Ren et al., 2026; Huang et al., 2026). An assistant must answer later questions from evidence distributed across past text, images, and sessions (Guo et al., 2026a; Chai et al., 2026; Liu et al., 2026a; Tang et al., 2026; Zhu et al., 2026; Li et al., 2026). In Figure 1, a user compares several chambray dress images and later refers to the one she likes most as "the third one." Answering which image she means requires matching that mention to the image order and distinguishing her preference from earlier candidate descriptions. The answer depends on whether an item was merely proposed, compared, or later preferred, in addition to its visible attributes.

Existing memory methods improve access to past interaction content through textual memory records, multimodal evidence stores, and structured retrieval interfaces (Zhong et al., 2024; Chhikara et al., 2025; Liu et al., 2026b; Feng et al., 2026; Yu et al., 2026; Guo et al., 2026b; Lewis et al., 2020; Borgeaud et al., 2022; Wu et al., 2022). One line of work textualizes histories into captions, summaries, facts, or structured records, making memory compact and searchable; Figure 1(a) shows this textualized-memory pathway. Another line keeps raw image-text records available for later use, preserving visual details that caption-only memory may lose; Figure 1(b) shows this raw-evidence route. These representations make memory records more compact, faithful, or connected. Yet retrieving relevant records does not by itself recover how the interaction organized them: which candidates were contrasted, what a later reference denoted, how a decision changed, and whether the resulting claim is supported (Shi et al., 2026; Ma et al., 2026a; Yang et al., 2026). We call this interaction-dependent configuration the *interaction state*. Without it, a system may retrieve the correct records yet bind an indirect reference to the wrong candidate, confuse an intermediate preference with the final decision, or answer beyond the available evidence.

The distinction is between recalling an item and recalling its role in the interaction. An image caption can describe a dress, but identifying it as the user's preferred option also requires linking the ordinal mention to that image and following the later preference. Even when retrieval supplies all relevant

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

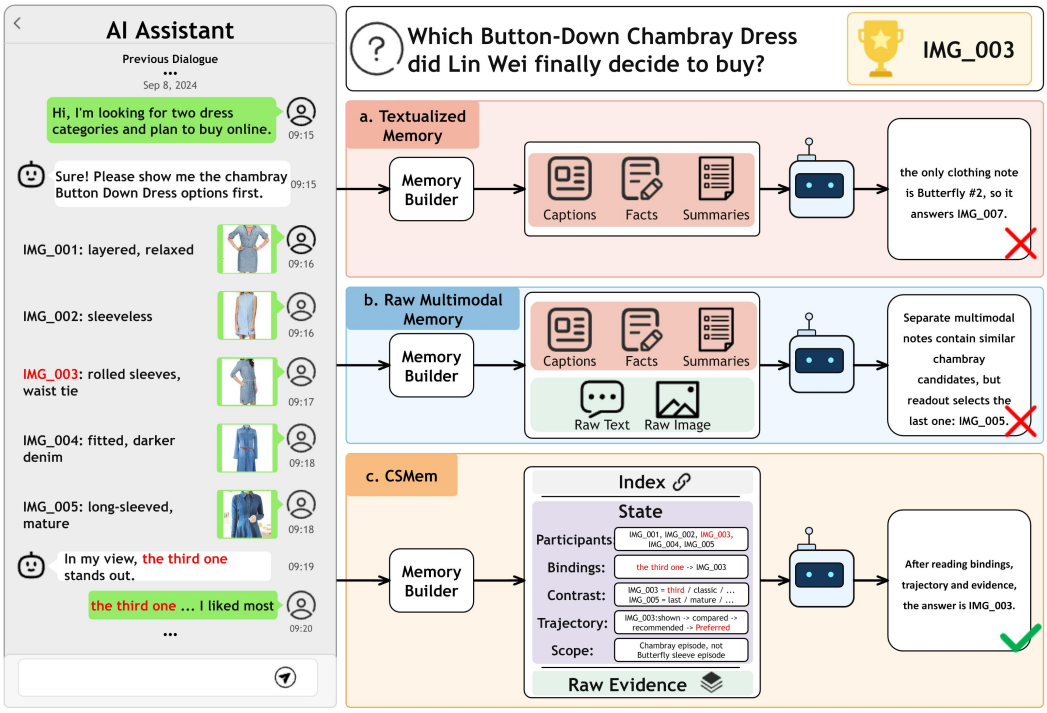


Figure 1: Motivation for storing interaction state. Identifying the dress the user favors requires linking "the third one" to the corresponding image and the later preference. (a) Textualized memory supplies captions, facts, or summaries. (b) Raw multimodal memory supplies the original image-text records, from which the binding and preference must be recovered. (c) CSMem stores these relations as Episode State and retains their source evidence for grounded answering.

turns, these relations still have to be recovered. CSMem records them during memory construction, while the episode context is available, and reuses them across later questions. In Figure 1, the binding resolves "the third one" and the trajectory links that candidate to the expressed preference; the original turns and images remain available to ground the answer.

Situation-model accounts of comprehension motivate this choice of representation: they describe understanding through entities and their configuration along dimensions such as time, space, causation, and intentionality (Zwaan & Radvansky, 1998). We use this perspective to organize answer-relevant information within a dialogue episode. Here, "cognitive state" refers to a task-oriented representation of the interaction established by the dialogue.

An *Episode State* makes five aspects of this organization explicit: which participants appear, which expressions refer to them, how candidates differ, how their roles change, and which claims the episode supports. We represent these aspects as participants, reference bindings (which mention refers to which entity), contrast dimensions (the attributes used to compare candidates), a state trajectory (how a role changes over the dialogue), and evidence scope (what the episode supports), yielding $S_i = (P_i, B_i, C_i, T_i, E_i)$. Participants provide the shared anchors: bindings resolve mentions to those anchors, contrasts distinguish alternatives, trajectories preserve decisions and revisions, and scope records what the episode can support. Together, these fields connect an answer candidate to both its role in the interaction and its source evidence, as shown in Figure 1(c).

CSMem organizes each episode into three linked parts. The *Episode Index* stores event, language, and visual cues for locating it; the *Episode State* records its interaction-dependent relations; and *Raw Evidence* retains its source text and images. At query time, matched index cues identify relevant episodes, the retrieved state organizes the answer context, and selected raw evidence grounds the answer. The resulting contribution is an episode-level interface that connects stored interaction relations to their use in multimodal answering.

Our contributions are:

- We introduce a source-linked Episode State that organizes participants, references, comparisons, role changes, and evidence scope within a multimodal dialogue interaction.
- We propose a training-free memory framework that connects Index, State, and Evidence for construction, retrieval, and grounded answering.
- We evaluate the combined interface on Mem-Gallery (Bei et al., 2026) and LoCoMo (Maharana et al., 2024), reaching 71.02% F1 / 54.30% EM and 56.02% F1, respectively, in a fixed controlled instantiation.

2 Related Work

Textualized Agent Memory. Generative Agents (Park et al., 2023), MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), A-MEM (Xu et al., 2026), and MemoryOS (Kang et al., 2025) store, update, and retrieve persistent information from interaction histories. For image-text dialogue, these systems commonly construct textual memories from captions or extracted facts, which improves compactness and searchability but can lose visual distinctions and episode-specific roles. CSMem retains the multimodal record while making those roles explicit in an Episode State.

Raw Multimodal Memory. Multimodal memory methods preserve visual records for later retrieval. AUGUSTUS (Jain et al., 2025), OmniMem (Liu et al., 2026b), and M2A (Feng et al., 2026; Ma et al., 2026b) attach compact metadata or semantic memories to raw multimodal content, allowing detailed evidence to be reopened when needed. However, a retrieved record does not by itself specify whether an item was a candidate, a rejected alternative, or the final choice in an episode. CSMem couples the retained record with participant bindings, contrasts, trajectory, and evidence scope so that the retrieved evidence can be interpreted in its interaction context.

Graph-Structured Retrieval and Memory. GraphRAG methods (Peng et al., 2025; Gutiérrez et al., 2024; Chen et al., 2026; Lin et al., 2026; He et al., 2026) organize evidence through entities and relations, while Neural Graph Memory (NGM) (Fisher, 2025) supports associative and temporal traversal over multimodal episodic records. These methods connect related records across a memory store. CSMem addresses a complementary structure formed within each episode: reference-to-participant bindings, candidate-specific contrasts, ordered role changes, and evidence scope. This episode-level state is linked to the original text and images rather than replacing them.

3 Method

3.1 Task and Overview

Given a long-term dialogue history $\mathcal{H} = (x_1, \ldots, x_T)$ and a later question q, the task is to produce an answer grounded in the historical text and images. Each turn x_t may contain text, images, image references, timestamps, and a session identifier. CSMem has two phases (Figure 2). Before queries arrive, it segments the history into episodes and constructs linked index, state, and evidence records. For each later query, it retrieves cue records from the index, gathers the states of their episodes, and uses the retrieved state to organize the answer context, resolve interaction-dependent references, and select supporting evidence.

An *episode* is a session-local interaction, such as a group of candidate images discussed before a decision. For episode i, we define the interaction record e_i and its associated raw evidence R_i as follows:

$$e_i = (s_i, I_i, y_i), \quad I_i = (t_{i,1}, \ldots, t_{i,n_i}), \quad R_i = (x_{t_{i,1}}, \ldots, x_{t_{i,n_i}}). \tag{1}$$

Here s_i is the session identifier, I_i lists its source turns in order, y_i is a coarse event label for the interaction, and R_i is the corresponding raw multimodal evidence. The event label is used as one search cue; it does not define an additional memory layer. The stored *Episode Memory* is

$$M_i = (K_i, S_i, R_i), \qquad \mathcal{M} = \{M_i\}_{i=1}^{N}. \tag{2}$$

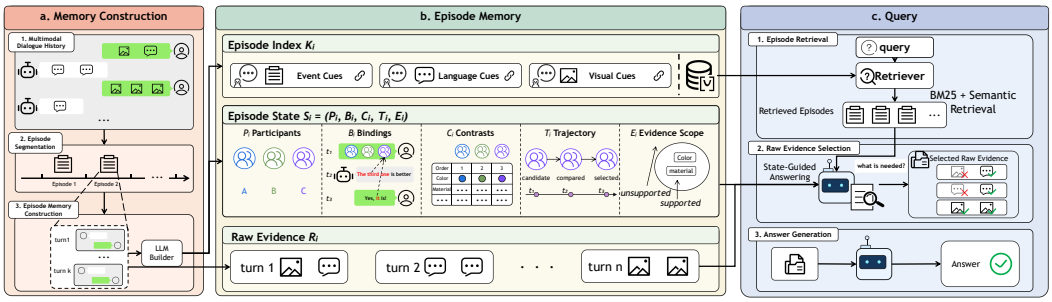


Figure 2: Overview of CSMem. (a) Construction segments the history into episodes and extracts their state records. (b) Each stored Episode Memory links an Episode Index K_i, an Episode State S_i, and Raw Evidence R_i. (c) At query time, matched index cues identify episodes; the stored state and selected raw evidence then assemble the context for answering. Sections 3.2–3.4 follow this construction-to-query workflow.

K_i is the Episode Index, S_i is the Episode State, and N is the number of episodes. Thus, e_i identifies an interaction in the input, M_i is its stored memory record, and S_i is the structured state within that record. The index locates the interaction, the state records its participants and relations, and the raw evidence preserves the source text and images used to ground answers.

3.2 Memory Construction

Episode segmentation. Construction begins by grouping turns within each session. A rule-based pass proposes boundaries using topic-shift expressions, new image groups, and decision or closure cues. The episode-memory builder can then refine these candidates to keep related comparisons, decisions, and corrections together. A coverage check requires every source turn to belong to exactly one resulting episode; if refinement fails this check, the rule-based segmentation is retained. Raw turns and images are preserved in R_i. Appendix C gives the boundary rules and refinement instructions.

State extraction. The LLM Builder runs five field-specific language-model prompts over the same episode payload, which contains its turn text, image identifiers, captions, and source metadata. Each prompt extracts one state field and returns JSON records for one component of

$$S_i = (P_i, B_i, C_i, T_i, E_i). \tag{3}$$

These components capture five aspects of the interaction: the participants, their referring expressions, their distinguishing attributes, their changing roles, and the scope of supported claims. The construction prompts and output schemas appear in Appendix J.

Participants. P_i records the people, objects, agents, and image-grounded entities involved in e_i. A participant can be grounded in text, an image, or both. Participant records retain episode and source identifiers, so similar visual candidates can be distinguished within the interaction. For image-grounded candidates, they also retain the image identifiers used by bindings and contrasts to refer to the same visual evidence. In Figure 1, the dress images are separate participants even when their captions are similar.

Reference bindings. Let L_i contain referring mentions together with their source turns. For single-target references, the binding is a partial map

$$B_i : L_i \rightharpoonup P_i. \tag{4}$$

It links a mention ℓ to participant p when supported by the interaction. Anchoring mentions to source turns distinguishes separate occurrences of "this one." An ordinal expression such as "the third one" can therefore point to the corresponding dress image.

Contrast dimensions. C_i records the dimensions used to distinguish participants in the episode:

$$C_i = \{(a, v_a) \mid a \in \mathcal{A}_i,\ v_a : P_i \rightharpoonup \mathcal{V}_a\}. \tag{5}$$

Here $\mathcal{A}_i$ is the set of recorded axes, such as color, material, order, location, or time, and $\mathcal{V}_a$ is the value domain for axis a. The partial map v_a associates a participant with its recorded value. For the dresses in Figure 1, sleeve style and presentation can distinguish candidates with otherwise similar descriptions.

State trajectory. T_i records changes in a participant's dialogue role, together with the turns that establish them:

$$T_i(p) = \big((\omega_j, t_j, r_j)\big)_{j=1}^{m_p}. \tag{6}$$

Here m_p is the number of recorded updates, ω_j is a role or state, t_j is its source turn, and r_j is the supporting source pointer. Roles include candidate, compared, recommended, selected, rejected, updated, and contradicted. Source-turn positions supply the temporal order. For a final-choice question, the relevant record is a supported selection or subsequent revision, rather than simply the latest mention of the item.

Evidence scope. Let Φ_i contain the topics or propositions recorded by the scope extractor for episode e_i. We write

$$E_i : \Phi_i \rightharpoonup \{\textsc{supported}, \textsc{unsupported}, \textsc{unknown}\}. \tag{7}$$

The three values distinguish claims backed by episode evidence, claims not supported by this episode, and unresolved claims. This is a partial, episode-local record: an absent entry does not imply that a proposition is false or absent from the entire history. For example, the user's stated preference may be supported even when an exact price remains unknown.

The five fields operationalize the situation-model motivation for this task. In the running example, P_i identifies the dresses, B_i resolves the ordinal mention, C_i distinguishes the candidates, T_i records the expressed preference, and E_i marks what the episode supports. Their shared episode and source identifiers allow these records to be used together.

Consolidation. Let $\widetilde{S}_i = \mathrm{Ext}(R_i)$ denote the five extractors' proposals. The stored state is

$$S_i = \mathrm{Consolidate}(\widetilde{S}_i). \tag{8}$$

Consolidation assembles the typed JSON records by episode, removes duplicates using field-specific identifiers, and filters reference targets against known participant and image identifiers. It retains source metadata for evidence lookup and records coverage and image-reference diagnostics. Appendix C gives the record keys used by these checks. The resulting state organizes the extracted relations while preserving access to their source evidence.

Index construction. The Episode Index $K_i = (K_i^{\mathrm{evt}}, K_i^{\mathrm{lang}}, K_i^{\mathrm{vis}})$ contains event, language, and visual cues. The cue extractor produces records linked to episode and source-turn identifiers; deterministic cues retain text keywords, image-caption attributes, dates, and event labels. Cue-dependent retrieval motivates retaining several ways to access the same interaction (Tulving & Thomson, 1973). The cue records form the searchable index, while S_i and R_i remain linked through their identifiers. Memory construction is performed before answering the later questions.

3.3 Episode Retrieval

Given question q, retrieval first ranks cue records from the Episode Index. Let $\mathcal{K}$ be the indexed cue set and $r_b(u; q)$, $r_v(u; q)$ the BM25 and semantic-vector ranks of cue u. Define

$$g(r) = \begin{cases} 1/(c + r), & \text{if rank } r \text{ is available,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The fused score is

$$\rho(u, q) = \lambda_b g(r_b(u; q)) + \lambda_v g(r_v(u; q)). \tag{10}$$

The default retrieval settings are $c = 60$, $\lambda_b = 0.55$, $\lambda_v = 0.45$, and $k = 10$. Let H_q be the top-k cue records and $\iota(u)$ the episode identifier attached to cue u. The retrieved episodes and their states are

$$A_q = \{\iota(u) \mid u \in H_q\}, \qquad S_q = \{S_i \mid i \in A_q\}. \tag{11}$$

Several matched cues can refer to the same episode, so episode identifiers are deduplicated before collecting state records. Here k counts cue records; the corresponding episode set supplies both state records and source pointers.

3.4 State-Guided Answering

Using the retrieved state. The retrieved state organizes answer construction in two ways: it provides interaction relations to the answer model, and its source pointers guide evidence selection. The context includes contrasts, trajectories, and supported claims, while participant and reference records link mentions to source turns and images. Together, the state records and their source links support reference resolution, comparison, and questions about the final state. A fixed set of task-specific instructions supplies question-specific answer rules; boundary checks also expose unsupported or unknown scope entries. The Mem-Gallery point labels are dataset-provided evaluation annotations; the labels themselves are not exposed to the answer model or serialized as memory evidence. Appendix C gives the dataset-specific rules.

Raw evidence selection. Raw evidence is exposed in priority order: source turns attached to matched cues, available immediately preceding and following turns in the same session, and then remaining turns from retrieved episodes. Duplicate turns are removed and an evidence budget limits this initial expansion. The last step adds further context from episodes already retrieved; we call this step *backfill*. For visual questions, reference and contrast information can further select image groups within those episodes. The answer prompt prioritizes direct cue evidence and instructs the model to use neighboring or backfilled evidence only when it supports the same event. Thus, selecting a neighboring turn and deciding whether it supports an answer are separate steps.

Answer generation. Let R_q denote the selected source text and images and τ_q the answer instructions. Source-based rules can directly return an exact image reference or date; other queries are answered by the multimodal generator G from q, the serialized state context S_q, R_q, and τ_q, followed by task-specific formatting and source-based postprocessing. Appendix C details these rules. For the dress example, the binding identifies the referenced image, the trajectory identifies the stated preference, and the selected dialogue turn and image provide its evidence. This completes the index–state–evidence pathway: retrieve the interaction, use its stored relations, and ground the answer in the original multimodal record.

4 Experiments

4.1 Experimental Setup

Benchmarks and data. We evaluate all 1,711 Mem-Gallery QA pairs and the 1,986 questions in LoCoMo-full10 (Bei et al., 2026; Maharana et al., 2024). Memory construction uses only source histories; gold answers, clue annotations, category labels, adversarial answers, and official summaries are withheld. Dataset details are in Appendix A.

Model configuration. CSMem is training-free. All methods use GPT-4o-mini for memory construction and Qwen3-VL-8B-Instruct (Bai et al., 2025) for answering under the same split and evaluation configuration. Baselines retain their released implementations and prompts; no method is fine-tuned on either benchmark. Experiments used eight NVIDIA GeForce RTX 4090 GPUs.

Baselines. Following the comparison in Figure 1, we group baselines into two families. The first family uses textualized or structured memory records, including MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), SimpleMem (Liu et al., 2026c), A-MEM (Xu et al., 2026), MemoryOS (Kang et al., 2025), and NGM (Fisher, 2025). The second family keeps raw multimodal evidence available during retrieval, including AUGUSTUS (Jain et al., 2025), OmniMem (Liu et al.,

Table 1: Controlled results grouped by the two baseline families in Figure 1. All model-mediated answers use Qwen3-VL-8B-Instruct (Bai et al., 2025); exact image-reference and date cases follow the shared source-based answer rules. All methods use GPT-4o-mini for memory construction. Bold and underline denote the best and second-best results per metric.

Method	LoCoMo						Mem-Gallery		
	MH	SH	Tmp	Open	Adv	Overall	F1	BLEU-1	EM
MemGPT (Packer et al., 2023)	0.2563	0.3864	0.2132	0.1173	0.3900	0.3277	0.4983	0.5006	0.3588
Mem0 (Chhikara et al., 2025)	0.2059	0.3374	0.0823	0.1581	0.8094	0.3746	0.5450	0.5622	0.3521
SimpleMem (Liu et al., 2026c)	0.2164	0.3438	0.1590	0.1323	0.8610	0.4018	0.5623	0.5305	0.4342
A-MEM (Xu et al., 2026)	**0.3298**	0.4629	0.2867	0.1356	0.8609	0.4891	<u>0.6642</u>	0.6143	<u>0.4880</u>
MemoryOS (Kang et al., 2025)	0.2973	0.3694	0.1166	<u>0.2386</u>	0.0291	0.2356	0.5258	0.5360	0.3488
NGM (Fisher, 2025)	0.0247	0.0430	0.0282	0.0560	**0.9551**	0.2436	0.5078	0.4669	0.3646
AUGUSTUS (Jain et al., 2025)	0.0942	0.2287	0.2451	0.1036	<u>0.8812</u>	0.3527	0.6350	0.5897	0.4711
OmniMem (Liu et al., 2026b)	0.3111	0.4368	0.2715	**0.2495**	0.8655	0.4794	0.6499	**0.6817**	0.4836
M2A (Feng et al., 2026)	0.3163	<u>0.5322</u>	<u>0.3522</u>	0.2022	0.7601	<u>0.5077</u>	0.6266	0.5776	0.4530
CSMem (Ours)	<u>0.3295</u>	**0.5847**	**0.4263**	0.1264	0.8498	**0.5602**	**0.7102**	<u>0.6659</u>	**0.5430**

2026b), and M2A (Feng et al., 2026). This grouping matches the main motivation: prior systems either retrieve textual or structured proxies, or retrieve original image-text records, while CSMem additionally exposes the Episode State induced by the interaction.

Evaluation protocol. All rows in Table 1 are single full runs on the same split and question set. Appendix C gives reproduction details and dataset-specific context construction.

Metrics. For Mem-Gallery, we report question-level F1, BLEU-1, and exact match (EM); for LoCoMo, category-level and overall F1. Definitions and category distributions are in Appendices A–B; higher is better.

4.2 Main Results

Table 1 summarizes the controlled comparison on LoCoMo (Maharana et al., 2024) and Mem-Gallery (Bei et al., 2026). Baselines are arranged by the memory families in Figure 1: textualized or structured memories, followed by raw multimodal memories.

Under the reproduced setting, CSMem obtains the best F1 and EM among the compared methods. It reaches 71.02% F1 and 54.30% EM, compared with 66.42% and 48.80% for A-MEM, the strongest baseline on these metrics. OmniMem remains stronger on BLEU-1, with 68.17% compared with 66.59% for CSMem, while CSMem's advantage is reflected in F1 and exact match. This pattern is consistent with the role of Episode State in questions about dialogue-grounded participants, decisions, and evidence scope.

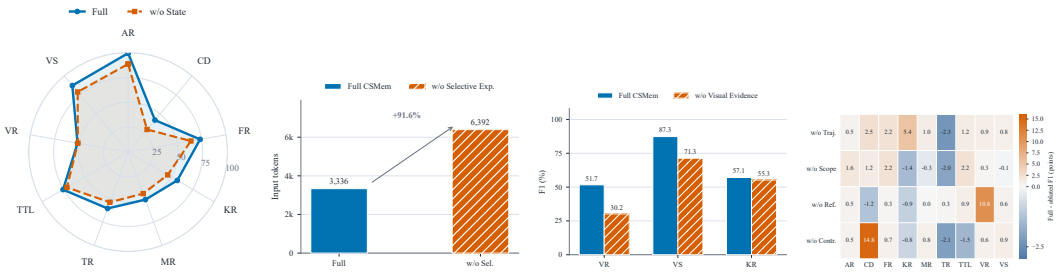
On LoCoMo, the annotation-free CSMem full10 run reaches 56.02% overall F1, compared with 50.77% for the strongest baseline, M2A. It also obtains 32.95% multi-hop, 58.47% single-hop, 42.63% temporal, 12.64% open-domain, and 84.98% adversarial F1. CSMem leads on single-hop, temporal, and overall F1, while A-MEM performs best on multi-hop, OmniMem on open-domain questions, and NGM on adversarial questions. This text-heavy diagnostic complements Mem-Gallery by testing long-range recall under a different modality profile; open-domain questions remain CSMem's main weakness.

4.3 Ablation Study

We use two ablation groups. The first group examines explicit Episode State with episode-level evidence expansion, selective evidence exposure, and raw image evidence. *w/o State* removes $S_i = (P_i, B_i, C_i, T_i, E_i)$ and the additional evidence expansion from retrieved episodes, answering from direct cue evidence and its images. *w/o Selective Exposure* keeps the state representation

Table 2: Ablation configurations and scores on Mem-Gallery. P, B, C, T, E denote participants, reference bindings, contrast dimensions, state trajectory, and evidence scope; Sel./Img. denote selective exposure/raw image evidence. Bold denotes the best score.

Variant	**State fields**					**Evidence**		**Score**		
	P	B	C	T	E	Sel.	Img.	F1	B1	EM
Full CSMem	✓	✓	✓	✓	✓	✓	✓	**0.7102**	**0.6659**	**0.5430**
w/o State	×	×	×	×	×	✓	✓	0.6387	0.5959	0.4927
w/o Selective Exp.	✓	✓	✓	✓	✓	×	✓	0.6731	0.6287	0.5108
w/o Visual Evidence	✓	✓	✓	✓	✓	✓	×	0.6599	0.6160	0.4944
w/o Trajectory	✓	✓	✓	×	✓	✓	✓	0.6988	0.6543	0.5330
w/o Scope	✓	✓	✓	✓	×	✓	✓	0.7030	0.6598	0.5365
w/o Reference	✓	×	✓	✓	✓	✓	✓	0.6962	0.6524	0.5295
w/o Contrast	✓	✓	×	✓	✓	✓	✓	0.7035	0.6615	0.5348



(a) Point-level F1 (b) Input tokens (c) Visual-evidence F1 (d) Field effects

Figure 3: Ablation diagnostics on Mem-Gallery. (a) Point-level F1 for Full CSMem and w/o State. (b) Average input tokens per question with and without selective exposure. (c) F1 on the three categories with the largest visual-evidence losses. (d) Full minus ablated F1 (percentage points) for individual state fields. Warmer heatmap cells indicate larger losses.

but reopens all retrieved evidence in the answer context. *w/o Visual Evidence* keeps the state and textual event context but removes memory-side image inputs. The first comparison therefore evaluates the combined state-and-expansion interface; the other two examine selective evidence exposure and memory-side image evidence.

The second group tests the internal fields of Episode State. Participants P_i remain as the anchor for state entries, while *w/o Trajectory*, *w/o Scope*, *w/o Reference*, and *w/o Contrast* ablate T_i, E_i, B_i, and C_i, respectively. For B_i and C_i, the corresponding use of the state field in evidence selection and answer construction is disabled together with the field. All variants use the same question set, answer model, and memory-construction model as the full system. Together, the two ablation groups examine the utility of the interface and the roles of its fields.

The architecture variants show distinct effects. Removing State and its episode-level evidence expansion causes the largest drop (F1 71.02%→63.87%, EM 54.30%→49.27%); the largest point changes occur on CD, KR, AR, FR, and VS (Figure 3(a)). Removing selective exposure lowers F1 to 67.31% and nearly doubles input tokens (3,336→6,392), while removing memory-side images lowers F1 to 65.99%, mainly on VR and VS.

The field variants localize the intended roles: w/o Reference mainly affects VR and w/o Contrast mainly affects CD. Trajectory has a smaller, category-dependent effect, while Scope is weakest and most diffuse.

4.4 Analysis by Question Category

We inspect official Mem-Gallery categories linked to Episode State. VR/VS test visual binding, KR/TR/MR test updated relations and order, CD/AR test evidence support, and TTL tests applying a learned concept to a query image. Table 3 reports the corresponding F1; FR is omitted because it mainly tests ordinary content recall.

Table 3: Mem-Gallery category F1 (%). State fields show the corresponding variables; *w/o* S removes Episode State and its evidence expansion. Bold and underline mark the highest and second-highest scores.

Point	State fields	N	CSMem	w/o S	M2A	A-MEM	AUG.	OmniMem
AR	E_i	184	**99.46**	88.59	93.40	94.80	93.05	<u>95.69</u>
CD	C_i, E_i	81	41.98	29.63	60.49	49.38	<u>62.96</u>	**82.72**
TTL	P_i, E_i	337	75.58	71.11	57.03	<u>77.89</u>	**82.29**	63.57
VR	B_i, E_i	174	**51.68**	<u>51.29</u>	48.08	48.43	40.66	36.87
VS	P_i, B_i, E_i	306	**87.33**	78.94	74.02	<u>82.03</u>	73.43	74.58
KR	T_i, E_i	81	**57.09**	45.98	44.90	41.92	30.49	<u>53.87</u>
TR	T_i	123	**60.39**	53.73	57.56	44.47	50.84	<u>59.73</u>
MR	P_i, C_i, T_i	206	<u>50.89</u>	44.45	49.15	48.68	43.53	**51.50**

Table 4: Mem-Gallery retrieval analysis (%). F1 uses all 1,711 QA pairs; Any, Rec., and All are computed on the 1,527-question clue-annotated subset.

Method	F1	Any	Rec.	All
CSMem	**71.02**	**92.73**	**81.59**	**67.26**
AUGUSTUS	63.50	78.45	62.59	46.50
M2A	62.66	65.09	50.06	39.29
SimpleMem	56.23	26.39	19.63	13.29
NGM	50.78	53.96	41.05	28.23

Across categories, CSMem improves over w/o S, with the largest gains on CD, KR, AR, VS, TR, and MR. It leads on AR, VR, VS, KR, and TR; OmniMem leads on CD/MR and AUGUSTUS on TTL. A qualitative example appears in Appendix G.

The category profile also separates different demands on memory. Relative to w/o S, the gains on KR and TR are 11.11 and 6.66 percentage points, respectively. The VS gain is 8.39 points, whereas the VR gain is only 0.39 points: the combined interface has a larger effect on finding the relevant remembered images than on visual reasoning in this comparison. These category-level differences describe the combined state-and-expansion interface; they do not isolate the contribution of any individual state field.

4.5 Evidence Retrieval Analysis

We evaluate retrieval against annotated clue turns. Table 4 reports answer F1 over all 1,711 QA pairs. The 184 AR questions have no positive clue turn, so Any@10, Recall@10, and All@10 are reported on the remaining 1,527 clue-annotated questions.

CSMem has the highest clue-turn coverage among baselines with turn-level traces (All@10: 67.26% vs. 46.50% AUGUSTUS, 39.29% M2A, and 28.23% NGM), indicating more complete evidence retrieval.

The gap between Any@10 and All@10 distinguishes finding some relevant evidence from recovering the complete annotated set. For CSMem, these scores differ by 25.47 percentage points. Its higher All@10 supports the completeness advantage over the compared retrievers, while the remaining gap shows that a successful clue match does not guarantee full coverage. This distinction matters when an answer must combine evidence from several turns.

5 Conclusion

We introduced Cognitive State Memory, a training-free framework that treats interaction state as a first-class memory object for multimodal long-term dialogue. By linking Episode State to retrieval cues and raw multimodal evidence, CSMem preserves how participants, references, contrasts, decisions, and evidence scope are organized within a remembered interaction. Experiments on Mem-Gallery and LoCoMo show gains on the main aggregate metrics over the strongest reproduced baselines. Future work will improve state extraction, reference binding, and evidence-scope calibration in longer and noisier multimodal memory settings.

AI USE STATEMENT

As components of the evaluated memory systems, GPT-4o-mini was used for memory construction and Qwen3-VL-8B-Instruct for model-mediated answer generation, as described in the Method and Experimental Setup sections. Separately, generative AI tools were used only to polish the manuscript's language. The authors reviewed the AI-assisted text and checked the numerical comparisons against the reported results. The authors take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Huacan Chai, Yukai Wang, Yingxuan Yang, Dan Peng, Yuanyi Song, Zhihui Fu, Weiwen Liu, Jianghao Lin, Jun Wang, and Weinan Zhang. Smmbench: A benchmark for source-distributed multimodal agent memory. *arXiv preprint arXiv:2605.15710*, 2026.

Xingyu Chen, Junxiu An, Jun Guo, and Li Wang. Page-rag: Evidence-grounded adaptive graph retrieval for long-document question answering. *arXiv preprint arXiv:2607.19301*, 2026.

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.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

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

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026a.

Weidong Guo, Dakai Wang, Zixuan Wang, Hui Liu, and Yu Xu. T-mem: Memory that anticipates, not archives. *arXiv preprint arXiv:2606.15405*, 2026b.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Xin He, Yili Wang, Wenqi Fan, Qing Li, Qinggang Zhang, Yi Chang, and Xin Wang. Hvm-graphrag: Holistic-view multimodal graph retrieval-augmented generation on complex document, 2026. URL https://arxiv.org/abs/2607.24861.

Zhengjun Huang, Wenxuan Liu, Zhoujin Tian, Wei Chen, Junle Chen, Yuqian Wu, Fangyuan Zhang, Qintian Guo, and Xiaofang Zhou. M^3Exam: Benchmarking multimodal memory for realistic user-agent interactions. *arXiv preprint arXiv:2606.07402*, 2026.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. Augustus: An llm-driven multimodal agent system with contextualized user memory. *arXiv preprint arXiv:2510.15261*, 2025.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

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

Ruizhe Li, Mingxuan Du, Benfeng Xu, and Zhendong Mao. Keep it inmind: Benchmarking the implicit-association blind spot in agent memory, 2026. URL https://arxiv.org/abs/2607.24368.

Jiashi Lin, Changhong Jiang, Xiangru Lin, Ruifei Zhang, Xinyi Zhu, Jiyao Liu, Cheng Tang, Ye Du, Shujian Gao, Junzhi Ning, et al. Evograph-r1: Self-evolving multimodal knowledge hypergraphs for agentic retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 756–765, 2026.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, et al. Worldmemarena: Evaluating multimodal agent memory through action-world interaction. *arXiv preprint arXiv:2605.29341*, 2026a.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, et al. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory. *arXiv preprint arXiv:2604.01007*, 2026b.

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

Enze Ma, Yufan Zhou, Wei-Chieh Huang, Jie Yang, Huanhuan Ma, Zixuan Wang, Chengze Li, Chunyu Miao, Philip S. Yu, and Zhen Wang. Memaudit: Auditing long-term agent memory via hidden user-state recovery, 2026a. URL https://arxiv.org/abs/2606.24595.

Xiaokang Ma, Yifan Sun, Zhihong Jin, Jie Gu, Yudong Luo, Shenyi Shao, Chu Tang, Jingmin Chen, and Li Pu. Reflectworld-mm: An entity-oriented multimodal memory system for open-ended video streams, 2026b. URL https://arxiv.org/abs/2607.09759.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52, 2025.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, et al. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-tma: Decoupling state-aware memory failures in long-term agent memory. *arXiv preprint arXiv:2607.01935*, 2026.

Yujin Tang, Chenming Shang, Ruize Xu, and Nikhil Singh. Dmv-bench: Diagnosing long-horizon multimodal agents' visual memory with incidental cue injection. *arXiv preprint arXiv:2606.27499*, 2026.

Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International conference on learning representations*, 2022.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Wei Yang, Bryce Kan, Shixuan Li, Li Li, Yuehan Qin, Jiate Li, Paul Bogdan, and Jesse Thomason. Ramem: Contextual reinstatement for long-term agentic memory. *arXiv preprint arXiv:2606.22844*, 2026.

Jing Yu, Yibo Zhao, Jiaming Zhang, and Xiang Li. Lazymem: Retrieve broadly, construct selectively for efficient long-term agent memory, 2026. URL https://arxiv.org/abs/2607.22690.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Shiping Zhu, Yibo Yang, Zhengyang Wang, Tiancheng Shen, Dandan Guo, and Ming-Hsuan Yang. H2hmem: A multimodal memory benchmark for agents in human-human interactions. *arXiv preprint arXiv:2606.09461*, 2026.

Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162, 1998.

APPENDIX OVERVIEW

This appendix provides dataset and evaluation details, implementation settings, additional analyses, and prompt templates. The following guide summarizes each section; the section titles and page numbers link to the corresponding material.

A. Dataset and Task Details p. 13
Summarizes the evaluated dataset copies, including session, turn, image, and question counts, and lists the Mem-Gallery point labels and LoCoMo categories.

B. Evaluation Metrics p. 14
Defines answer normalization, F1, BLEU-1, and exact match, together with category-specific LoCoMo scoring and the clue-turn retrieval metrics Any@k, Recall@k, and All@k.

C. Implementation and Configuration p. 15
Describes baseline reproduction, model and retrieval settings, episode segmentation and consolidation, and the protocols for assembling evidence and producing answers.

D. Efficiency Analysis p. 16
Compares average end-to-end query latency and answer accuracy for CSMem and the reproduced baselines on Mem-Gallery and LoCoMo.

E. Retrieval Hyperparameter Sensitivity p. 16
Examines how lexical/vector fusion weights and the retrieval budget affect Mem-Gallery answer scores and retrieval diagnostics while other components remain fixed.

F. Error Analysis p. 17
Breaks down residual exact-match errors by clue coverage in the final answer context, a heuristic error taxonomy, and Mem-Gallery question category.

G. Qualitative Case Study p. 18
Compares the recovered memory views and answers of A-MEM, OmniMem, and CSMem on a visual-search question involving foxglove images and nearby distractors.

H. Limitations p. 18
Discusses memory-construction overhead, the scope of the ablation evidence, and the need for a comparison with a generic structured memory of matched capacity.

I. Future Work p. 19
Outlines evaluation on additional multimodal memory benchmarks and refinement of the state representation, particularly trajectory and evidence scope.

J. Prompts p. 19
Provides the five state-field extraction prompts and their JSON schemas, along with the answer-generation prompt template.

A DATASET AND TASK DETAILS

Dataset scale. Table 5 summarizes the local copies used in our experiments. Mem-Gallery provides multimodal multi-session dialogues, image assets, question images, point labels, and clue-turn annotations. LoCoMo provides very long multi-session conversations with QA categories and evidence turn identifiers; image URLs and generated image captions are included, while image files are not locally released by the dataset.

Mem-Gallery point labels. Table 6 lists the Mem-Gallery point labels and the local distribution. These dataset-provided labels define the point-level analyses. They are evaluation annotations: the labels themselves are not supplied as prompt inputs or included in memory evidence; the answer protocol follows the fixed benchmark-compatible rules described below.

Table 5: Dataset scale in the local evaluation copy. Mem-Gallery contains 20 scenarios, while LoCoMo-full10 contains 10 conversations. LoCoMo image count refers to image-URL/image-turn entries.

Dataset	Sess.	Turns	Images	QA
Mem-Gallery	240	3,962	1,490	1,711
LoCoMo-full10	272	5,882	910	1,986

Table 6: Mem-Gallery point labels and local distribution over 1,711 QA pairs.

Point	Name	QA	Q-Img.
FR	Factual retrieval	219	0
VS	Visual search	306	88
TTL	Test-time learning	337	333
TR	Temporal reasoning	123	0
VR	Visual reasoning	174	54
MR	Multi-entity reasoning	206	0
KR	Knowledge resolution	81	2
CD	Conflict detection	81	10
AR	Answer refusal	184	0

LoCoMo categories. For LoCoMo, we follow the category mapping in the evaluation code. Table 7 lists the five categories used in Table 1.

Table 7: LoCoMo category mapping and local QA counts.

Code	Name	QA
C1	Multi-hop	282
C4	Single-hop	841
C2	Temporal	321
C3	Open-domain	96
C5	Adversarial	446

B Evaluation Metrics

Mem-Gallery answer metrics. Mem-Gallery scores are question-level averages over all evaluated QA pairs. Before scoring, both prediction and gold answer are normalized by lowercasing, removing articles, removing "and", replacing most punctuation with spaces, and preserving decimals and underscores so that image identifiers such as `D8:IMG_003` remain meaningful. Exact match is

$$\mathrm{EM}(\hat{a}, a) = \mathbb{I}[\mathrm{norm}(\hat{a}) = \mathrm{norm}(a)]. \tag{12}$$

Token F1 is computed from normalized token overlap:

$$\mathrm{F1} = \frac{2PR}{P+R}, \quad P = \frac{|Y_{\hat{a}} \cap Y_a|}{|Y_{\hat{a}}|}, \quad R = \frac{|Y_{\hat{a}} \cap Y_a|}{|Y_a|}. \tag{13}$$

When Porter stemming is available, the scorer applies it before token overlap. BLEU-1 is unigram precision with a brevity penalty:

$$\text{BLEU-1} = \mathrm{BP} \cdot \frac{\sum_w \min(c_{\hat{a}}(w), c_a(w))}{\sum_w c_{\hat{a}}(w)}. \tag{14}$$

LoCoMo answer metrics. LoCoMo uses category-specific F1. C1 multi-hop questions split comma-separated answer parts and score the set of parts. C2 temporal and C4 single-hop questions use normalized token F1. C3 open-domain questions take the maximum F1 over semicolon-separated acceptable answer parts. C5 adversarial questions score whether the prediction correctly

abstains with a no-information style answer. Overall LoCoMo F1 is the question-level average over the evaluated QA pairs.

Retrieval metrics. For retrieval diagnostics on Mem-Gallery, let C_q be the annotated clue-turn set. We normalize the recorded retrieval trace to source-turn identifiers, retain the first k identifiers in trace order, and form the set $\widehat{C}_q^{(k)}$. For CSMem, this diagnostic uses the recorded cue-turn trace; for baselines, it uses their retrieved identifiers mapped to dialogue turns. Thus, k is a turn-trace cutoff for these metrics, and the evaluated set is distinct from the final answer context after evidence expansion. We report:

$$\text{Any@}\, k = \mathbb{I}[C_q \cap \widehat{C}_q^{(k)} \neq \emptyset], \tag{15}$$

$$\text{Recall@}\, k = \frac{|C_q \cap \widehat{C}_q^{(k)}|}{|C_q|}, \qquad \text{All@}\, k = \mathbb{I}[C_q \subseteq \widehat{C}_q^{(k)}]. \tag{16}$$

These metrics are computed only on questions with clue annotations. The 184 answer-refusal questions have the reference answer "Not mentioned" and no positive clue turn, so they are outside this retrieval matching analysis.

C Implementation and Configuration

Baseline reproduction. For the main comparison in Table 1, we reproduce every baseline using its open-source implementation under the same shared configuration: GPT-4o-mini for memory construction and Qwen3-VL-8B-Instruct for model-mediated answer generation across all methods. Exact image-reference and date cases follow the shared source-based answer rules. We retain the prompt templates provided by each method; when a required template is unavailable, we use the corresponding benchmark prompt.

CSMem configuration. CSMem is a training-free memory framework. The experiments use off-the-shelf models for memory construction, retrieval support, and model-mediated answer generation. Qwen3-VL-8B-Instruct is used as the controlled answer model; exact image-reference and date cases follow the shared source-based answer rules. Memory is constructed from the source histories. Gold answers, annotated clue turns, adversarial answers, and official summaries are not provided as answer evidence. The two datasets use the answering protocols specified below.

Table 8: Main experimental configuration.

Parameter	Value
Construction model	GPT-4o-mini
Answer model	Qwen3-VL-8B-Instruct
Embedding model	BAAI/bge-large-en-v1.5
Retrieval fusion	$\lambda_b = 0.55, \lambda_v = 0.45, c = 60$
Retrieval budget	top-$k = 10$ cue records
Hardware	8 NVIDIA GeForce RTX 4090 GPUs
Fine-tuning	none

Implementation mapping. The paper uses conceptual names for readability. In the implementation, the Episode Index is stored as cue records, the Episode State as episode-level state rows, and Raw Evidence in the raw backend. The five state fields correspond to participants, reference bindings, contrast dimensions, state trajectory, and evidence scope.

Episode boundaries and consolidation. The initial segmenter groups turns by session. Once a span has at least eight turns, an incoming image or topic-shift cue starts a new span; a decision or closure cue closes the current span. An image accompanied by a topic-shift cue can also start a span after at least three turns when the preceding turn has no image. Eight is a boundary threshold, not a hard maximum episode length. Refinement receives the session turns and candidate spans and is instructed to preserve related candidate groups, comparisons, decisions, and corrections. Refined

spans are accepted only if they cover every source turn exactly once. Consolidation removes duplicate participant records by episode, participant, and source turn; reference records by episode, source turn, span, and target; contrasts by episode and axis; trajectories by episode and participant; and scope records by episode. Invalid reference targets are filtered against known participant and image identifiers.

Answering protocol. For Mem-Gallery, the point labels are not supplied to the model, used as prompt content, or included in memory evidence. Answer handling follows the fixed benchmark-compatible operation and answer-format rules described below. LoCoMo uses the question and retrieved history without the official C1–C5 category labels, gold answers, or annotated evidence. In both cases, we first gather the state records belonging to the retrieved episodes. Context assembly includes event summaries, contrasts, trajectories, and supported scope entries; boundary-checking additionally exposes the recorded unsupported or unknown scope. Reference and participant records also support evidence lookup.

Evidence expansion and answer handling. The initial raw-evidence expansion retains up to 14 unique turns, prioritizing direct cue turns, immediate same-session neighbors, and then remaining turns from retrieved episodes. Visual-group selection can add evidence from those retrieved episodes, so the final context may contain additional source turns. The answer prompt prioritizes direct evidence and restricts the use of neighboring evidence to the same event. Exact image-reference or date questions can return a source-backed answer directly; other questions use the answer model, with task-specific answer formatting and source-based postprocessing.

D Efficiency Analysis

We report average end-to-end per-query latency. Each measurement covers retrieval, context construction, the model API request, and answer handling.

Table 9: Mem-Gallery answer accuracy and average per-query latency. F1 and EM are percentages.

Method	F1	EM	Mean (s)
CSMem	71.02	54.30	4.39
A-MEM	66.42	48.80	1.12
OmniMem	64.99	48.36	13.73
AUGUSTUS	63.50	47.11	2.02
M2A	62.66	45.30	1.44
SimpleMem	56.23	43.42	14.05
Mem0	54.50	35.21	0.77
MemoryOS	52.58	34.88	0.92
NGM	50.78	36.46	1.10
MemGPT	49.83	35.88	11.77

Table 10: LoCoMo Overall F1 and average per-query latency. F1 is a percentage.

Method	F1	Mean (s)
CSMem	56.02	3.80
M2A	50.77	2.28
A-MEM	48.91	0.88
OmniMem	47.94	4.77
SimpleMem	40.18	18.19
Mem0	37.46	0.67
AUGUSTUS	35.27	0.55
MemGPT	32.77	3.14
NGM	24.36	0.46
MemoryOS	23.56	0.85

E Retrieval Hyperparameter Sensitivity

We evaluate two retrieval hyperparameters on the full Mem-Gallery set: the lexical/vector fusion weight and the retrieval budget. These runs keep the Episode State, answer model, prompts, and memory construction fixed. The auxiliary "Cue Rec." and "Event Rec." columns retain the names used by the evaluation archive; they are sweep diagnostics and are separate from the clue-turn metrics reported in the main paper.

The fusion sweep shows that the answer score is stable over mixed lexical/vector settings. Pure BM25 and pure vector retrieval are weaker, while the default 0.55/0.45 setting is close to the best observed score and keeps the two retrieval signals balanced. The top-k sweep shows the expected

Table 11: Sensitivity to lexical/vector fusion on Mem-Gallery. Scores are percentages; 0.55/0.45 is the default used in the main experiments.

BM25/Vec.	F1	EM	Cue Rec.	Event Rec.
1.00/0.00	68.85	52.43	72.12	65.68
0.75/0.25	70.50	53.89	81.04	75.34
0.55/0.45	71.02	54.30	81.59	76.10
0.25/0.75	71.29	54.76	82.43	76.26
0.00/1.00	70.07	53.77	77.67	72.93

Table 12: Sensitivity to retrieval budget on Mem-Gallery with BM25/vector weights fixed at 0.55/0.45.

Top-k	F1	EM	Cue Rec.	Event Rec.
1	59.21	45.88	32.45	10.34
5	70.11	53.48	66.65	50.86
10	71.02	54.30	81.59	76.10
20	70.62	54.24	90.56	90.52
30	69.35	53.42	93.45	94.56

coverage/noise tradeoff. Top-$k = 1$ lacks evidence coverage; top-$k = 20$ and top-$k = 30$ increase clue recall but do not improve final answer quality, indicating that adding more evidence can introduce distracting context.

F Error Analysis

We analyze the full CSMem Mem-Gallery run reported in the main paper. We use exact-match errors because EM is the strictest reported metric and exposes failures that token F1 can partially credit. Evidence coverage is computed from annotated clue turns in the final answer context, and error types are derived from the question label and the observable prediction/gold-answer form.

Table 13: Evidence coverage of residual exact-match errors. Numbers are percentages over the 782 EM errors of the full CSMem Mem-Gallery run.

Final context	QA	Share	Interpretation
All clues	576	73.7	answer mismatch
Partial clues	165	21.1	partial evidence
No annotated clue	40	5.1	no clue retrieved
No positive clue	1	0.1	answer refusal
Total EM errors	782	100.0	–

Table 13 shows that 73.7% of EM errors occur despite all annotated clue turns being present in the final context, while 5.1% contain none of those turns. Evidence coverage alone therefore does not ensure an exact answer; most remaining errors require examining how the available context is interpreted and the answer is formed.

Tables 14–15 break down these residual errors. The largest heuristic groups are incomplete semantic answers (32.3%, mostly in MR and FR) and visual discrimination or wrong image binding (33.7% combined, in TTL, VR, and VS). Temporal and latest-state errors account for 17.7%. At the point level, MR and FR have the most errors; 85.8% of MR errors occur with all annotated clue turns present, motivating closer inspection of multi-entity answer completeness. The full system makes one EM error on AR under the evaluated answering protocol.

Table 14: Heuristic residual taxonomy for the 576 EM errors where all annotated clues are present in the final answer context. Shares are percentages within these complete-evidence errors.

Error type	QA	Share	Typical source
Incomplete semantic answer	186	32.3	MR, FR
Visual interpretation	132	22.9	TTL, VR
Temporal/latest-state	102	17.7	TR, KR, MR
Image/entity binding	62	10.8	VS, VR
False abstention	56	9.7	TTL, KR, VS
Conflict judgment	34	5.9	CD
Other substitution	4	0.7	mixed

Table 15: Point-level residual hotspots. "All clues" is the percentage of EM errors in that point whose final context contains all annotated clue turns.

Point	EM err.	All clues	Main residual issue
MR	190	85.8	multi-entity completeness
FR	129	57.4	long-answer realization
TTL	122	83.6	query-image transfer
VR	96	72.9	visual counting
VS	81	76.5	similar-image binding
TR	64	60.9	temporal ordering
KR	52	59.6	latest-state selection
CD	47	74.5	conflict polarity
AR	1	0.0	rare boundary miss

G Qualitative Case Study

All three methods retrieve the episode containing the relevant evidence, but their answers differ when constructing the final answer. CSMem exactly matches the gold answer, while A-MEM and Omni-iMem produce non-exact answers despite preserving relevant evidence in their recovered memory views. Figure 4 illustrates the underlying mechanism. The answer is the set of images sharing the same species as the query image within the same interaction, rather than a nearest-neighbor image match. The raw dialogue establishes five foxglove images (`D8:2--D8:6`), then shifts to mallow (`D8:8--D8:10`) and later to a Beagle. Textualized memory preserves the foxglove facts but does not mark the species boundary as an interaction state; raw multimodal memory retains visual evidence but exposes session-level distractors from nearby events. CSMem separates these by representing the remembered interaction through bindings, contrasts, and trajectory: `this one`/`this angle` are bound to specific foxglove images, mallow turns are recorded as a distinct sequence, and the answer draws only from raw evidence linked to the foxglove state.

H Limitations

First, memory construction in CSMem costs more than storing flat captions or summaries, because it runs five field-specific extractors per episode on top of the model calls shared with textualized baselines.

Second, the ablations show unequal effects across components: reference binding and visual evidence yield clearer gains, while trajectory and evidence scope show more diffuse effects. The architecture-level w/o State variant removes both state records and additional evidence expansion from retrieved episodes, so its overall drop reflects their combined contribution. Finer-grained validation of trajectory and scope representations remains needed.

Third, a matched-capacity generic structured memory or free-form episode summary would provide a useful comparison for the particular organization of Episode State.



Figure 4: A Mem-Gallery visual-search case from `Baking_Dessert_Daily_Life_Skill`. The query image shows a foxglove and asks for all conversation photos of the same plant species. The gold set is `D8:IMG_001--D8:IMG_005`. A-MEM retrieves the relevant foxglove text records but over-includes three mallow images in its output; OmniMem's recovered input view contains foxglove, mallow, and Beagle evidence, yet its output misses two foxglove images and includes a Beagle distractor. CSMem uses the Episode State and raw evidence, separating the foxglove sequence from later mallow and Beagle evidence, and returns the exact gold set.

I Future Work

We plan to evaluate CSMem on additional long-term multimodal assistant benchmarks to further test the generality of the Episode Index–Episode State–Raw Evidence interface. We also aim to refine the state structure so that each field—particularly trajectory and evidence scope—contributes more reliably across question categories.

J Prompts

Prompt 1: Participant binding

```
Given the event payload, identify text/image participants in the event.

Return strict JSON:
```

```
{
  "participant_binding": [
    {
      "participant_id": "...",
      "event_id": "...",
      "modality": "text|image",
      "source_turn": "...",
      "role": "dialogue_context|visual_evidence|candidate|reference|
    decision_evidence",
      "image_id": "...",
      "image_path": "...",
      "caption": "...",
      "text_span": "..."
    }
  ]
}

Payload:
{payload}
```

Prompt 2: Reference binding

```
Given the event payload, resolve dialogue references such as "this one", "first
    one", "burgundy one", "that item".

Return strict JSON:

{
  "reference_binding": [
    {
      "event_id": "...",
      "source_turn": "...",
      "text_span": "...",
      "refers_to": "image_id or participant_id",
      "grounding_cues": ["..."]
    }
  ]
}

Payload:
{payload}
```

Prompt 3: Contrast extraction

```
Given the event payload, identify contrast axes among participants in the same
    event.

Return strict JSON:

{
  "contrast_axis": [
    {
      "event_id": "...",
      "axis": "color|ordinal|breed|category|material|location|action|time|
    visual_attributes",
      "values": {"participant_or_image_id": "value"}
    }
  ]
}

Payload:
{payload}
```

Prompt 4: State trajectory extraction

```
Given the event payload, extract dialogue-triggered role state trajectories.

Allowed states: observed, candidate, compared, recommended, selected, rejected,
    updated, contradicted, unknown.

Return strict JSON:

{
  "state_trajectory": [
    {
      "event_id": "...",
      "participant_id": "...",
      "states": [
        {"state": "...", "at": "turn_id", "trigger": "..."}
      ]
    }
  ]
}

Payload:
{payload}
```

Prompt 5: Scope-boundary extraction

```
Given the event payload, extract evidence scope boundaries.

Return strict JSON:

{
  "scope_boundary": [
    {
      "event_id": "...",
      "supported": ["topics or facts supported by the event"],
      "not_mentioned": ["topics explicitly outside evidence if any"],
      "unknown": ["topics that should not be guessed"],
      "source_turns": ["..."]
    }
  ]
}

Payload:
{payload}
```

Prompt 6: Answer prompt

```
You are an AI assistant evaluated on multimodal long-term conversational memory.
Answer concisely using only the provided memory evidence.
Read the direct cue evidence before neighboring or backfilled evidence. Use
    neighboring evidence only to
complete the same event, never to introduce a second topic. Before answering,
    check that the subject,
relation, object, and time requested by the question are supported by the same
    evidence frame.
Use "Not mentioned" only when the relevant topic/entity is absent from the
    retrieved evidence.
For yes/no questions, return "Yes." or "No." unless the topic is absent.
For before/after questions, return only "Before." or "After." when supported.
For list questions, return a compact comma-separated list grounded in the same
    event and remove duplicates.
Return only the answer content, without "answer:" or explanations.

Memory evidence:
{context}

Task policy:
{task_policy}

Question: {question}
```