BEYOND RANKING: MAKING LONG-TERM MULTIMODAL MEMORY ADDRESSABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term multimodal memory is useful only if a future query can find the evidence it needs. Yet storing the right observation does not make it easy to retrieve, especially when evidence is visual, surrounded by similar memories, or distributed across multiple memory units. We call this problem the *addressability gap*. Across three long-term multimodal memory benchmarks, we study this gap through three retrieval challenges: finding image-bearing evidence, distinguishing the required observation from similar memories, and recovering the complete evidence set. We introduce CACE, an addressability-oriented framework with two components. At construction time, Contrastive Addressability Encoding (CAE) describes each observation in the context of visually similar memories, making its distinguishing details searchable while preserving the original observation as evidence. At retrieval time, Competitive Evidence Assignment (CEA) preserves individual query tokens and competitively assigns each token's matching credit across candidate memories, allowing different parts of the query to recruit different textual or visual evidence under a fixed retrieval budget. Experiments on Mem-Eye, SMMBench, and Mem-Gallery show substantial improvements in evidence retrieval across all three challenges, with corresponding gains in downstream answer quality across multiple reader models. These results suggest that long-term multimodal memory should be designed not only to store relevant experience, but to make that experience addressable by future queries.

1 INTRODUCTION

Multimodal agents are moving from isolated interactions toward persistent assistants that accumulate experience across days, months, and evolving tasks (Park et al., 2023; Zhong et al., 2024). Their histories contain not only dialogue, but also images, screenshots, documents, and other artifacts whose relevance may emerge only much later (Maharana et al., 2024; Wang et al., 2024; Xie et al., 2024). As these histories grow beyond the context window, long-term multimodal memory becomes essential(Chhikara et al., 2025; Xu et al., 2026; Wang & Chen, 2025; Wu et al., 2024). It stores past observations and retrieves relevant evidence when new questions arise (Packer et al., 2023; Guu et al., 2020; Gutiérrez et al., 2024).

Recent multimodal memory systems have developed increasingly sophisticated mechanisms for storing, organizing, and retrieving such experience (Feng et al., 2026; Liu et al., 2026; Lu et al., 2026). Yet recent benchmarks reveal a persistent problem: *storing the right evidence does not guarantee that a future query can locate it*. A system may preserve an answer-bearing image but fail to retrieve it from a textual question; it may retrieve the right type of observation but confuse it with a similar one; or it may recover some of the required evidence while missing the rest (Guo et al., 2026; Chai et al., 2026; Bei et al., 2026). These failures arise before the answering model can reason over the evidence and therefore cannot be resolved by better downstream reasoning alone.

We study this problem through three retrieval challenges. **First, evidence must be findable.** Important information such as numbers, labels, counts, and object states may appear only in an image and remain poorly represented in the searchable text. A natural-language question can therefore miss the relevant observation altogether. **Second, evidence must be distinguishable.** Long-term memory often contains repeated or visually similar observations that share the same topic but differ in the exact detail required by the question. Locating the right semantic neighborhood is therefore not

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.

enough; retrieval must identify the particular observation being asked about. **Third, evidence must be jointly retrievable.** Many questions depend on several memory units, so independently ranking relevant memories does not ensure that the complete evidence set fits within a limited retrieval budget (Xiong et al., 2020).

In this work, we refer to the common requirement behind these challenges as *evidence addressability*: a stored observation is addressable when a future query can identify and retrieve it from memory. Addressability is especially challenging in long-term multimodal memory because future questions are unknown during memory construction. The system must therefore encode cues that make an observation searchable before knowing which of its properties will later matter. At retrieval time, it must use those cues to recover the required evidence among many related memories and within a given retrieval budget. This perspective shifts the focus from whether relevant experience is merely stored to whether a future query has an effective way to reach it.

We introduce **CACE**, a framework designed to improve evidence addressability in long-term multimodal memory. At construction time, **Contrastive Addressability Encoding (CAE)** places each image in the context of visually similar memories and generates an indexing description that captures both its content and the attributes that distinguish it from its neighbors. The original observation remains unchanged and is returned as evidence; the generated description serves only as a searchable representation. At retrieval time, **Competitive Evidence Assignment (CEA)** preserves individual query tokens rather than collapsing the query into a single vector. For each token, candidate memories compete for matching credit based on their textual and visual evidence. Different parts of the query can therefore recruit different memories, and the resulting assignments are aggregated to select evidence under a fixed retrieval budget.

Our key contributions are threefold. **First**, we formulate *evidence addressability* as a new perspective on three recurring failures in long-term multimodal memory: finding visual evidence, distinguishing it from similar observations, and retrieving complete evidence sets. **Second**, we introduce CACE, which combines contrastive addressability encoding at construction time with competitive token-level evidence assignment at retrieval time. **Third**, we evaluate evidence retrieval separately from downstream answering across various multimodal memory benchmarks, showing how improvements in memory construction and retrieval affect both evidence access and end-task performance.

2 Related Work

Multimodal memory systems explore consolidation, structured organization, and reliability-aware memory management (Feng et al., 2026; Liu et al., 2026; Lu et al., 2026). However, recent benchmarks reveal persistent difficulties in retaining fine-grained visual details and retrieving evidence distributed across observations (Guo et al., 2026; Chai et al., 2026; Bei et al., 2026). Research on retrieval accessibility and distinctive image captioning further highlights the importance of making stored items retrievable and distinguishable (Azzopardi & Vinay, 2008; Wang et al., 2020a). Building on these findings, we study evidence addressability in long-term multimodal memory. Due to space limitations, we provide a detailed discussion in **Appendix A**.

3 The Addressability Gap in Long-Term Multimodal Memory

The *addressability gap* arises when the evidence needed to answer a question is stored in memory but cannot be reliably identified and retrieved. In multimodal histories, relevant details may appear only in images, similar observations may differ in the specific detail being queried, and a single question may require evidence from several memory units. We therefore examine three corresponding retrieval challenges under a fixed budget: finding image-bearing evidence, distinguishing the required observations from similar memories, and recovering complete evidence sets.

3.1 Setup and Measurements

Experimental setup. We conduct our study on MemEye, SMMBench, and Mem-Gallery (Guo et al., 2026; Chai et al., 2026; Bei et al., 2026), excluding SMMBench's function-call questions because their answer format is incompatible with our evaluation protocol. A memory unit is a

history entry with its native text, images, and provenance; retrieval is restricted to the question's corpus or scenario. The evaluation sets contain 371 questions on MemEye, 1,769 on SMMBench, and 1,711 on Mem-Gallery. Mem-Gallery spans 20 scenarios; 1,527 questions have gold evidence and 184 do not. Evidence retrieval metrics use only questions with nonempty gold evidence sets.

We diagnose two baselines on all three benchmarks. SRAG-V is MemEye's dense multimodal retriever (`semantic_rag_multimodal`). Its score equally weights text and visual similarity. The text term is the cosine similarity between all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) embeddings of the question and the unit's native text; the visual term is the maximum question–image cosine similarity over the unit's images, using SigLIP 2 (`base-patch16-384`) (Tschannen et al., 2025). SimpleMem-V is the visual configuration of Omni-SimpleMem (Liu et al., 2026). In the evaluated adapter, it indexes dataset-provided text and captions and returns the linked source units with their associated images. All diagnostics use a fixed retrieval budget of $K = 10$ memory units. Downstream answer quality is evaluated separately in Section 5.

Measurements. Let $\mathcal{M} = \{m_i\}_{i=1}^{N}$ be a memory collection, where N is the number of memory units and $m_i = (x_i, I_i, p_i)$ contains native text x_i, an image set I_i, and provenance p_i. Given a question q, a retriever returns the top-K memory units, denoted by $R_K(q)$. The benchmark provides a gold evidence set G_q, and we define $G_q^{\text{vis}} \subseteq G_q$ as the units that contain at least one image. Let $\mathcal{Q}$ denote the evaluated questions with nonempty gold evidence sets. Retrieval and gold evidence sets are represented by memory-unit indices.

We evaluate evidence access at two levels. **Image-bearing evidence recall** measures the fraction of image-bearing gold units retrieved:

$$\text{ImgRecall@}K = \frac{\sum_{q\in\mathcal{Q}} |R_K(q) \cap G_q^{\text{vis}}|}{\sum_{q\in\mathcal{Q}} |G_q^{\text{vis}}|}. \tag{1}$$

Counts are pooled across questions rather than averaged per question. Each image-bearing gold unit is counted once per question, regardless of how many images it contains. Questions without image-bearing gold evidence contribute zero to both the numerator and denominator.

Complete retrieval measures the fraction of questions for which the full gold evidence set is retrieved:

$$\text{Complete@}K = \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \mathbf{1}[G_q \subseteq R_K(q)]\,. \tag{2}$$

The indicator $\mathbf{1}[\cdot]$ counts a question as complete only when every gold unit, including text-only evidence, is retrieved; questions receive equal weights. $\text{ImgRecall@}K$ measures how much image-bearing evidence is retrieved, whereas $\text{Complete@}K$ requires the entire gold evidence set for a question. We report both metrics at $K = 10$.

3.2 Gap I: Limited Retrieval of Image-Bearing Evidence

Answering questions over long-term multimodal memory requires retrieving the stored observations that contain the relevant visual evidence. The first addressability gap arises when these observations are available in memory but are not retrieved by a natural-language query within the retrieval budget.

Table 1: **Image-bearing evidence retrieval.**

Retriever	ImgRecall @10 (%) ↑	Zero-evidence (%) ↓
MemEye		
SRAG-V	53.8	21.8
SimpleMem-V	51.2	19.7
SMMBench		
SRAG-V	45.7	23.2
SimpleMem-V	27.4	44.1
Mem-Gallery		
SRAG-V	63.4	14.5
SimpleMem-V	55.9	17.9

We characterize this gap with a retrieval budget of ten memory units. As shown in Table 1, SRAG-V retrieves 45.7–63.4% of image-bearing gold units, while SimpleMem-V retrieves 27.4–55.9%. Thus, neither direct multimodal similarity nor caption-mediated retrieval consistently recovers the stored visual evidence. At the question level, 14.5–23.2% of queries with SRAG-V and 17.9–44.1% with SimpleMem-V retrieve no gold evidence of either modality.

Table 2: **Similar-neighbor retrieval.**

Retriever	Complete@10 (%) ↑
MemEye 341 (91.9%)	
SRAG-V	43.4
SimpleMem-V	43.1
SMMBench 922 (52.1%)	
SRAG-V	15.2
SimpleMem-V	9.2
Mem-Gallery 976 (63.9%)	
SRAG-V	41.6
SimpleMem-V	38.2

Table 3: **Multi-evidence retrieval.**

Retriever	Complete@10 (%) ↑	Gold-unit miss rate (%) ↓
MemEye 252/371 (67.9%)		
SRAG-V	45.0	46.2
SimpleMem-V	45.0	48.8
SMMBench 1,736/1,769 (98.1%)		
SRAG-V	27.4	47.5
SimpleMem-V	18.1	62.4
Mem-Gallery 948/1,527 (62.1%)		
SRAG-V	52.8	37.6
SimpleMem-V	50.2	41.4

These results identify an evidence-access bottleneck in the evaluated retrieval setting: a substantial fraction of the gold visual evidence is excluded from the observations supplied to the answering model. The gap therefore arises before downstream reasoning and motivates improving the connection between natural-language queries and stored visual observations.

3.3 Gap II: Distinguishing Evidence Among Similar Memories

The second gap concerns distinguishing the required evidence from similar memories. Similar memories may share the same scene or object while differing in the number, label, or state needed to answer a question. A retriever that matches their shared content but misses these distinctions can select a related observation instead of the required evidence, consuming the limited retrieval budget. We examine how common such similar alternatives are and how well retrievers recover complete evidence sets in their presence. We identify questions with at least one gold unit that has a similar neighbor in the same corpus or scenario, defined by cosine similarity above 0.8 between pooled original-text and image representations.

Table 2 shows that questions whose gold evidence has similar neighbors account for 52.1–91.9% of each benchmark, yet Complete@10 is only 9.2–43.4% across the two baselines. The prevalence of similar memories, together with limited retrieval completeness, suggests that similar alternatives may interfere with retrieval, making it harder to recover the evidence required by the question.

3.4 Gap III: Incomplete Retrieval of Multiple Evidence Units

The third gap concerns retrieving the complete evidence set for a question. When a question requires several memory units, retrieving individual relevant observations is not enough: the required evidence must be recovered together within the retrieval budget. We first measure how common multi-evidence questions are and how often retrievers recover complete evidence sets.

Table 3 shows that 62.1–98.1% of questions require multiple gold units. Across all evaluated questions, Complete@10 is only 27.4–52.8% for SRAG-V and 18.1–50.2% for SimpleMem-V. The two baselines miss 37.6–62.4% of gold units. Thus, multi-evidence requirements are common, while complete retrieval remains limited.

To examine how completeness varies with evidence requirements, we group questions by the number of image-bearing gold units and measure whether all such units are retrieved. Questions without image-bearing gold evidence are excluded.

As shown in Figure 1, for both baselines, visual Complete@10 decreases monotonically with visual-gold count on MemEye and Mem-Gallery. SMMBench is non-monotonic, but questions requiring at least four visual units have the lowest completion rate on all three benchmarks. Recovering the full visual evidence set is therefore particularly difficult for questions with many required observations.

Together, these results highlight three requirements for long-term multimodal memory: find the evidence, distinguish it from similar memories, and retrieve the complete evidence set.

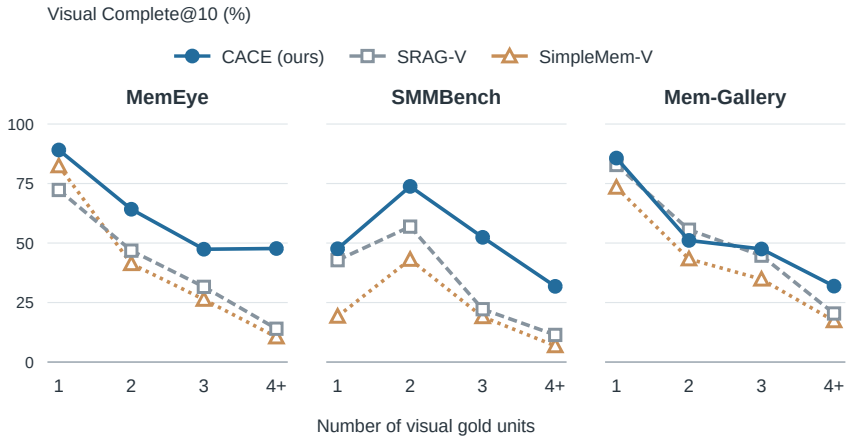


Figure 1: **Visual-evidence completeness by evidence count.** Visual Complete@10 measures whether all image-bearing gold units are retrieved under a common Top-10 budget.

4 CACE: Addressability Encoding and Evidence Assignment

The three gaps in Section 3 motivate two complementary design choices in CACE. During memory construction, visual details must become searchable without losing the attributes that distinguish similar observations. *Contrastive Addressability Encoding (CAE)* therefore generates indexing descriptions in the context of neighboring images. During retrieval, a question may require evidence from several memories, whereas independent scoring can repeatedly reward matches to the same part of the question. *Competitive Evidence Assignment (CEA)* therefore compares candidate memories for each query token before aggregating their scores, allowing different query details to favor different observations.

The construction stage uses only the memory collection, without access to evaluation questions, answer options, or evidence annotations. At retrieval time, CACE returns at most K memory units, whose original text and images are passed to the answering model.

4.1 Contrastive Addressability Encoding (CAE)

CAE constructs an index description for each image in the context of visually similar memories. The goal is to preserve its content while specifying attributes that distinguish it from otherwise similar observations.

Find visual alternatives. Let I_i denote the image set stored in memory m_i, and let $J_i \in I_i$ be the target image. A frozen visual encoder $v(\cdot)$ maps each image to a vector. We select up to r target images from other memories with the highest cosine similarity to J_i, subject to a minimum similarity of τ. Their memory-unit indices form

$$\mathcal{N}_i = \mathrm{Top}_r\{j \neq i : \cos(v(J_i), v(J_j)) \geq \tau\}. \tag{3}$$

These neighbors provide the comparison context for identifying attributes that distinguish the target image from similar memories. Neighbor selection uses only stored observations and is independent of future questions.

Describe the target in context. A multimodal writer receives the target image followed by its selected neighbors and generates an indexing description d_i. The instruction asks the writer to preserve the target's content while specifying attributes that distinguish it from the alternatives, such as objects, numbers, counts, states, and labels. Each target is processed independently; if no neighbor satisfies the threshold, it is described without comparison images. The description-generation prompt is provided in **Appendix B.1**.

Construct a searchable card. After text preprocessing, we concatenate the memory's native text x_i with the generated description:

$$c_i = x_i \,\|\, d_i. \tag{4}$$

Here $\|$ denotes text concatenation. A frozen text encoder h embeds each card as $h(c_i)$ for dense retrieval. The card serves only as an index: each entry links back to the original memory, and only the original text and images are delivered as evidence.

4.2 Competitive Evidence Assignment (CEA)

CEA first retrieves a candidate frontier and then performs fine-grained multimodal matching within it. Unlike independent late interaction, which scores each candidate in isolation, CEA compares candidates for each query token before aggregating their scores.

Candidate retrieval. We combine textual and visual similarity for coarse retrieval. The text encoder h embeds the question q and each index card c_i. The visual encoder v and its paired text encoder v_{txt} map stored images and the question into a shared space:

$$g_i = (1-w)\cos\big(h(q), h(c_i)\big) + w \max_{I \in I_i} \cos\big(v_{\text{txt}}(q), v(I)\big), \tag{5}$$

where $w \in [0,1]$ controls the relative weight of visual similarity. The visual term is zero for text-only memories. We retain the highest-scoring candidates:

$$\mathcal{F}(q) = \text{Top}_{F,i}\, g_i, \tag{6}$$

where $\mathcal{F}(q)$ contains the indices of up to F highest-scoring candidates and F is the candidate budget.

Fine-grained evidence matching. A frozen multimodal multi-vector encoder represents the question as $Q = \{\mathbf{q}_t\}_{t=1}^{T}$ and the original content of candidate i as an evidence bag $E_i = \{\mathbf{e}_{ip}\}_{p=1}^{P_i}$. Here T is the number of query-token vectors and P_i is the number of evidence vectors in candidate i. Each evidence vector corresponds to an original text token or image patch.

For each query token t and candidate i, we take its strongest nonnegative match over the candidate's evidence:

$$S_{ti} = \max_{p \in \{1,\dots,P_i\}} [\langle \mathbf{q}_t, \mathbf{e}_{ip} \rangle]_+, \qquad [z]_+ = \max(z, 0). \tag{7}$$

Thus, S_{ti} measures how strongly candidate i matches query token t, and $S \in \mathbb{R}^{T \times |\mathcal{F}(q)|}$ collects these token–candidate matches.

Competitive assignment. Standard late interaction scores each candidate independently as $s_i^{\text{ind}} = \sum_{t=1}^{T} S_{ti}$. Several memories can therefore rank highly by matching the same part of the question. CEA instead makes candidates compete for each query token: a stronger match receives a larger share of that token's matching credit.

$$A_{ti} = \frac{\exp(S_{ti}/\epsilon)}{\sum_{j \in \mathcal{F}(q)} \exp(S_{tj}/\epsilon)}, \qquad \epsilon = \text{std}(S). \tag{8}$$

For each token t, the weights A_{ti} sum to one across candidates. A candidate's weight thus depends on how well it matches the token relative to the other memories, not just on its own match strength. Different query tokens can favor different memories. The temperature ϵ controls how strongly the weights favor better matches; we compute it as the standard deviation of all scores in S.

We combine this relative allocation with the original match strength to compute $s_i = \sum_{t=1}^{T} A_{ti} S_{ti}$. Here, S_{ti} measures absolute token–candidate correspondence, while A_{ti} downweights matches that are less competitive relative to other candidates. The final retrieval set is $R_K(q) = \text{Top}_{K, i \in \mathcal{F}(q)}\, s_i$, where K is the final retrieval budget. This token-wise comparison allows memories supporting different query details to contribute to the final ranking, rather than rewarding only repeated matches to the same part of the question. It can therefore help recover evidence distributed across memories, although it does not guarantee a complete evidence set.

5 Experiments

We conduct experiments aiming to investigate three research questions: **(1)** Does CACE mitigate the three addressability gaps identified in Section 3? **(2)** Does improved evidence access translate into better downstream answering? and **(3)** How do CAE and CEA affect the retrieval performance?

5.1 Evaluation Setup

Benchmarks and metrics. We evaluate on MemEye, SMMBench, and Mem-Gallery (Guo et al., 2026; Chai et al., 2026; Bei et al., 2026). These benchmarks cover similar visual observations, evidence distributed across sources, and long multimodal conversational histories. Evidence retrieval is evaluated using ImgRecall@10 and Complete@10 as defined in Section 3.1. Downstream MemEye scores use the project's coordinate-balanced exact-match protocol across answer-option rotations. SMMBench uses exact match, while Mem-Gallery follows its benchmark judging protocol.

Baselines and answering models. We compare against four baselines from the MemEye evaluation benchmark, covering retrieval-based, summarization-based, and agentic memory systems (Guo et al., 2026). **SRAG-V** (Guo et al., 2026) is the retrieval-based baseline. It preserves the original text and images and retrieves ten memory units by combining their textual and visual similarity to the question. We use it as our primary baseline to assess improvements over direct retrieval from unmodified memories. **SimpleMem-V** (Liu et al., 2026) represents the summarization family: an LLM compresses the dialogue history into summaries that are searched at answer time, and the original images are attached to the retrieved summaries rather than being encoded. **M2A** (Feng et al., 2026) and **MMA** (Lu et al., 2026) are agentic memory systems, which build and query their own memory instead of searching a fixed index: M2A runs a multimodal reasoning-and-acting loop over a semantic memory bank, while MMA retrieves with confidence weights that combine similarity with the source and recency of each memory. We evaluate downstream answering with GPT-4.1 nano (Achiam et al., 2023), Gemini 2.5 Flash-Lite (Comanici et al., 2025), and Qwen3.7-Flash (Yang et al., 2025). For brevity, we refer to the latter two as Gemini and Qwen in the text; Table 7 uses their full names. Within every cell the compared systems share the answering model, so differences reflect the memories delivered rather than the reader.

Implementation. A single configuration is used for all three benchmarks and all three answering models; no value is tuned per dataset. Descriptions are written once by GPT-4.1 nano and then frozen. We use SigLIP 2 (`siglip2-base-patch16-384`) (Tschannen et al., 2025) for visual neighbor mining and cross-modal similarity in coarse retrieval. Neighbor mining keeps at most $r = 3$ neighbors with cosine similarity at least $\tau = 0.6$. Each card retains characters of native text and description, and is embedded by the MiniLM text encoder `all-MiniLM-L6-v2` (Wang et al., 2020b) under its 256-token window. We fix $w = 0.5$ in Equation 5 across all datasets, following MemEye's dense multimodal retriever and giving equal weight to text and visual similarity without dataset-specific tuning. Coarse retrieval keeps a frontier of $F = 30$ candidates, from which CEA selects $K = 10$ using `colqwen2-v1.0` representations of the original text and images. For Mem-Gallery, we follow the benchmark's LLM-as-a-judge evaluation protocol (Bei et al., 2026), using GPT-5 mini (Singh et al., 2025) as the judge.

5.2 Does CACE Mitigate the Three Addressability Gaps?

We first test whether CACE can mitigate the three retrieval failures identified in Section 3: inaccessible visual evidence, ambiguity among similar memories, and incomplete multi-evidence retrieval. We compare CACE with SRAG-V and SimpleMem-V using the protocol in Section 3.1 and a ten-memory delivery budget. Table 4 reports the three primary retrieval metrics. The full five-metric comparison is provided in **Appendix C**.

Accessing visual evidence. CACE improves ImgRecall@10 over SRAG-V by 14.6, 8.4, and 1.0 percentage points on MemEye, SMMBench, and Mem-Gallery, respectively. The corresponding gains over SimpleMem-V are 17.2, 26.7, and 8.5 points. Zero-evidence rates also decrease on all three benchmarks, falling from 21.8% to 8.6% on MemEye relative to SRAG-V. These results show that CACE retrieves more of the stored visual evidence and leaves fewer questions without any gold evidence. They support our design of making visual details searchable through indexing descriptions while retaining the original observations for fine-grained matching.

Distinguishing similar memories. CACE achieves the highest Neighbor Complete@10 on all three benchmarks, improving over SRAG-V by 21.1, 4.3, and 6.8 percentage points, respectively. On MemEye, the completion rate rises from 43.4% to 64.5%, showing that more questions receive the required evidence even when similar alternatives are present. This improvement supports CAE's contrastive description design, which makes the attributes distinguishing an observation from similar memories explicit and searchable.

Table 4: **Evidence retrieval (%, ↑).**

Method	ImgRecall @10	Neighbor Complete@10	Complete @10
MemEye			
SRAG-V	53.8	43.4	45.0
SimpleMem-V	51.2	43.1	45.0
CACE	**68.4**	**64.5**	**65.8**
SMMBench			
SRAG-V	45.7	15.2	27.4
SimpleMem-V	27.4	9.2	18.1
CACE	**54.1**	**19.5**	**32.2**
Mem-Gallery			
SRAG-V	63.4	41.6	52.8
SimpleMem-V	55.9	38.2	50.2
CACE	**64.4**	**48.4**	**58.7**

Table 5: **MemEye ablation.**

Method	EM ↑
w/o CAE	0.641
w/o CEA	0.675
CACE	**0.695**

Table 6: **Time cost.**

Method	Total time (min)
SRAG-V	9.7
CACE	11.7
MMA	39.9
SimpleMem-V	43.1
M2A	138.0

Recovering complete evidence sets. CACE achieves the highest Complete@10 on all three benchmarks. On Mem-Gallery, Complete@10 increases from 52.8% to 58.7%, despite only a 1.0-point improvement in ImgRecall@10. The benefit therefore extends beyond the pooled amount of visual evidence retrieved: more questions receive their entire evidence set within the same budget. This outcome is consistent with CEA's token-wise comparison, which allows memories supporting different parts of a question to contribute to the final ranking.

Effect of evidence multiplicity. Figure 1 examines whether these gains extend to questions requiring multiple visual observations. CACE leads in Visual Complete@10 in 11 of 12 benchmark–count groups, with gains of 11.5–33.7 percentage points over SRAG-V for questions requiring at least four visual gold units. The exception is the two-unit Mem-Gallery group, where it trails SRAG-V by 4.5 points. Overall, the results support the combined design of CACE: contrastive indexing makes specific visual observations easier to locate, while candidate-aware token matching helps recover evidence distributed across memories.

5.3 Does Better Evidence Access Improve Answering?

With the answering model fixed within each comparison, CACE achieves the highest reported score in eight of nine benchmark–reader combinations (Table 7). Reader-average scores improve over SRAG-V by 5.4% on MemEye, 7.2% on SMMBench, and 3.0% on Mem-Gallery, using each benchmark's own scoring scale. Better evidence access therefore translates into improved answering across distinct tasks and reader families.

The comparison also spans different memory designs. With the same reader, CACE outperforms both agentic baselines in every benchmark–reader combination and SimpleMem-V in eight of nine. More elaborate memory organization therefore does not by itself ensure better answers. The results instead support treating access to the required observations as a distinct design objective.

The results also suggest that stronger readers can better capitalize on improved evidence access. Gemini and Qwen gain on all three benchmarks; on MemEye, their relative improvements over SRAG-V are 9.4% and 6.3%, respectively. Their corresponding gains on SMMBench are 9.8% and 7.4%, compared with 4.7% for GPT-4.1 nano. This pattern suggests that reader capability can help realize the value of improved retrieval, although the size of that benefit also varies by benchmark. Making specific visual details and distributed evidence available creates an opportunity for capable readers to interpret and combine them, making retrieval and reader improvements complementary.

5.4 How Do CAE and CEA Contribute?

We examine the construction and retrieval stages through controlled substitutions. The w/o CAE variant replaces CAE with SRAG-V's memory construction while retaining CEA. The w/o CEA variant retains CAE but replaces the retrieval stage with SRAG-V's retrieval. On MemEye, these substitutions reduce exact match from 0.695 to 0.641 and 0.675, respectively (Table 5). Both sub-

Table 7: **Overall performance comparison of five memory systems across three benchmarks and three LLM backbones.** Reader Avg. is the arithmetic mean over the three readers within a benchmark, and is never averaged across benchmarks.

MemEye (coordinate-balanced exact match)

Method	GPT-4.1 nano	Gemini 2.5 Flash-Lite	Qwen3.7-Flash	Reader Avg.
CACE (ours)	0.512	**0.695**	**0.776**	**0.661**
SRAG-V	0.516	0.635	0.730	0.627
SimpleMem-V	**0.523**	0.474	0.563	0.520
MMA	0.434	0.498	0.546	0.493
M2A	0.391	0.338	0.531	0.420

Mem-Gallery (LLM judge)

Method	GPT-4.1 nano	Gemini 2.5 Flash-Lite	Qwen3.7-Flash	Reader Avg.
CACE (ours)	**0.696**	**0.765**	**0.773**	**0.745**
SRAG-V	0.670	0.742	0.757	0.723
SimpleMem-V	0.610	0.678	0.702	0.663
MMA	0.403	0.426	0.458	0.429
M2A	0.360	0.262	0.456	0.359

SMMBench (exact match)

Method	GPT-4.1 nano	Gemini 2.5 Flash-Lite	Qwen3.7-Flash	Reader Avg.
CACE (ours)	**0.557**	**0.629**	**0.639**	**0.608**
SRAG-V	0.532	0.573	0.595	0.567
SimpleMem-V	0.514	0.518	0.552	0.528
MMA	0.479	0.486	0.501	0.489
M2A	0.353	0.316	0.323	0.331

stitutions also lower answer quality on the other two benchmarks, as reported in **Appendix D**. The results support the complementary roles of our two stages: CAE provides searchable descriptions of distinguishing visual details, while the CEA-based retrieval stage compares fine-grained support across candidates. Combining them yields better answers than retaining either stage with its SRAG-V counterpart.

We also evaluate total execution time, including construction, retrieval, and answering, on Mem-Gallery with Gemini 2.5 Flash-Lite (Table 6). CACE improves answering over SRAG-V with 21% more total execution time and outperforms the summarization-based and agentic baselines at lower cost. The additional execution time is 2.0 minutes across 1,711 questions, with construction included in the measurement. In comparison, the other baselines require 39.9–138.0 minutes. Thus, the answer-quality gains do not require the runtime of the more elaborate memory pipelines evaluated here. The full results are reported in **Appendix E**.

6 Conclusion

This work identifies *evidence addressability* as a central requirement for long-term multimodal memory. Across three benchmarks, we show that evidence can remain difficult to retrieve even when it is preserved: visual observations are missed, similar memories obscure the required instance, and relevant units are retrieved without completing the evidence set. CACE tackles this gap through contrastive descriptions that make distinguishing details searchable and competitive token-level matching that connects different parts of a question to evidence across memories. The resulting improvements in retrieval and downstream answering support a shift in memory design: greater storage and organizational complexity should not be mistaken for better access to experience. Memory must make the required evidence identifiable and addressable.

AI USE STATEMENT

We used generative AI tools to improve the language, organization, and clarity of this manuscript, summarize related work, and assist in interpreting and discussing the experimental results. The authors reviewed the AI-assisted text and analytical suggestions, checked numerical comparisons against the reported results, and ensured that the final descriptions accurately reflected the methods and findings. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript, including all AI-assisted content.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Leif Azzopardi and Vishwa Vinay. Accessibility in information retrieval. In *European Conference on Information Retrieval*, pp. 482–489. Springer, 2008.

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.

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.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 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.

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*, 2026.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

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*, 2026.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. Mma: Multimodal memory agent. *arXiv preprint arXiv:2602.16493*, 2026.

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.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1070–1079. IEEE, 2017.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2020a.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Group-based distinctive image captioning with memory attention. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 5020–5028, 2021.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020b.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–442, 2024.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, et al. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*, 2020.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

CAE Description Prompt

The FIRST image is a memory; the other image(s) are its NEAR-DUPLICATES in memory. Write ONE description for the FIRST image that (1) lists EVERY object and attribute so it can be FOUND by a text search, AND (2) states the specific values, numbers, counts, states, or labels that make THIS one DISTINCT from the near-duplicates.

Figure 2: Prompt used for Contrastive Addressability Encoding (CAE).

A Related Work

Multimodal agent memory. Existing multimodal memory systems explore different strategies for representing, organizing, and maintaining accumulated experience. M2A separates an immutable raw-message store from editable semantic memories, retaining original observations alongside consolidated knowledge (Feng et al., 2026). Omni-SimpleMem organizes multimodal inputs into summary-based memory units linked to raw content, combining selective ingestion, graph-augmented search, and progressive content expansion (Liu et al., 2026). MMA incorporates source credibility, temporal decay, and consensus to manage the reliability of stored information (Lu et al., 2026). These approaches emphasize complementary aspects of memory management, including abstraction, structured organization, and reliability.

Multimodal memory benchmarks. Recent benchmarks reveal persistent limitations in these designs. MemEye finds that textual abstraction can discard fine-grained visual details, while retaining original images does not resolve the difficulty of selecting valid visual states as observations change over time (Guo et al., 2026). SMMBench shows that native multimodal access alone is insufficient: systems still struggle to retrieve, align, and combine evidence distributed across independent sources (Chai et al., 2026). Mem-Gallery reports that simple multimodal retrieval can outperform more elaborate memory architectures. It also finds that increasing retrieval coverage does not consistently improve answer quality, as additional results can introduce irrelevant information (Bei et al., 2026). Building on these findings, we examine which required observations are retrieved and whether the evidence set is complete under a fixed retrieval budget.

Addressability and distinctive descriptions. Accessibility in information retrieval concerns how readily documents can be reached through a retrieval system (Azzopardi & Vinay, 2008). Distinctive image captioning studies why an accurate description may nevertheless fail to distinguish an image from similar alternatives (Vedantam et al., 2017). Compare and Reweight uses similar-image sets to encourage distinctive captions and evaluates their retrieval utility (Wang et al., 2020a; 2021). CAE brings this perspective to memory construction, conditioning indexing descriptions on neighboring observations so that instance-specific visual details become searchable before future questions are known. These descriptions guide retrieval, while the original observations remain the evidence used for answering.

B Implementation and Evaluation Details

B.1 CAE Description Prompt

For the description-generation step in Section 4.1, the multimodal writer receives the target image first, followed by the selected visually similar images. Figure 2 shows the prompt.

C Evidence Retrieval Results

Table 8 extends the compact comparison in Table 4 with Zero@10 and Gold Miss@10.

Table 8: **Full evidence retrieval results across the three addressability gaps.** SMV denotes SimpleMem-V. All rates are percentages, and bold marks the best result for each benchmark and metric. Neighbor Complete@10 uses questions whose gold evidence has similar neighbors. Complete@10 and Gold Miss@10 use all questions with nonempty gold evidence.

Benchmark	Method	ImgRecall @10 (%) ↑	Zero @10 (%) ↓	Neighbor Complete@10 (%) ↑	Complete @10 (%) ↑	Gold Miss @10 (%) ↓
MemEye	SRAG-V	53.8	21.8	43.4	45.0	46.2
	SMV	51.2	19.7	43.1	45.0	48.8
	CACE	**68.4**	**8.6**	**64.5**	**65.8**	**31.6**
SMMBench	SRAG-V	45.7	23.2	15.2	27.4	47.5
	SMV	27.4	44.1	9.2	18.1	62.4
	CACE	**54.1**	**18.4**	**19.5**	**32.2**	**41.5**
Mem-Gallery	SRAG-V	63.4	14.5	41.6	52.8	37.6
	SMV	55.9	17.9	38.2	50.2	41.4
	CACE	**64.4**	**11.8**	**48.4**	**58.7**	**35.7**

Table 9: **Ablation study of CAE and CEA.**

Method	MemEye	SMMBench	Mem-Gallery
Ours w/o CAE	0.641	0.607	0.754
Ours w/o CEA	0.675	0.608	0.752
Ours	0.695	0.629	0.765

D How Do CAE and CEA Contribute?

We ablate the two components of CACE to examine their roles in memory construction and retrieval. **w/o CAE** replaces CAE with SRAG-V's memory construction while retaining CEA. **w/o CEA** replaces CEA with SRAG-V's retrieval while retaining CAE. All variants use the same memory collection, retrieval budget, and evaluation protocol.

Table 9 shows that replacing either CAE or CEA with its SRAG-V counterpart lowers answer quality on all three benchmarks. The two substitutions support complementary roles for searchable distinguishing details at construction time and fine-grained evidence selection at retrieval time, with both contributing to downstream answering.

E Time Cost

We measure total execution time on all 1,711 Mem-Gallery questions across 20 scenarios, using Gemini 2.5 Flash-Lite as the answering model. The reported time includes memory construction, retrieval, and answering, with relative cost normalized to SRAG-V.

Table 10 shows that CACE requires 11.7 minutes, compared with 9.7 minutes for SRAG-V, adding only 2.0 minutes or 21% overhead. In contrast, MMA, SimpleMem-V, and M2A require 39.9, 43.1, and 138.0 minutes, respectively, yet all obtain lower answer scores than CACE in this setting. CACE achieves the highest score of 0.765, improving over SRAG-V's 0.742 without a comparable increase in runtime.

These results challenge the assumption that more elaborate memory management necessarily produces more useful memory. Additional summarization, organization, or agentic processing can increase computational cost without resolving the central problem identified in the introduction: stored evidence may still be difficult for a future question to locate. In this evaluation, the substantially higher runtime of these systems does not translate into better answers. This favors a more targeted design: make distinguishing evidence searchable and improve its selection at retrieval time, rather than relying on increasingly complex memory-management pipelines to overcome an evidence-access bottleneck.

Table 10: **Time cost on Mem-Gallery with Gemini 2.5 Flash-Lite.** All methods answer the same 1,711 questions, and reported time covers memory construction, retrieval and answering. Relative cost is normalised to SRAG-V.

Method	Total time (min)	Relative cost
SRAG-V	9.7	1.00×
CACE (ours)	11.7	1.21×
MMA	39.9	4.11×
SimpleMem-V	43.1	4.44×
M2A	138.0	14.23×