CoReMem: Context-Aware Memory Construction and Query-Aware Retrieval for Multimodal Conversations

Anonymous authors
Paper under double-blind review

Abstract

Long-horizon multimodal conversational agents must recover relevant visual evidence from extended interaction histories. Existing memory systems typically convert images to text, potentially omitting visual details, or retrieve complete images using global embeddings, which can underrepresent small but relevant regions. Their retrieval procedures also commonly rely on a fixed top-k budget, which may omit evidence needed for multi-hop reasoning. We present CoReMem, a multimodal memory framework for context-aware construction and query-aware retrieval. During construction, a vision-language model uses only the dialogue observed so far to select potentially useful regions, SAM3 produces the corresponding masks, and validated regions become independently indexed subimage documents linked to their parent turns. During retrieval, hybrid text–image ranking first retrieves an initial ranked set of documents drawn from at most k dialogue turns. A VLM checks whether this initial context is sufficient for the user query. If not, CoReMem constructs visual queries conditioned on the query and images from the retrieved turns. It marks retrieved subimages as candidate query regions and uses DINOv3-based image-to-image retrieval to expand the context. The sufficiency check is repeated after each expansion to determine whether further evidence is needed. On MemEye with Qwen3.6-27B, CoReMem achieves LLM-as-a-Judge score of 59.8, outperforming the strongest evaluated baseline by 8.6. It also improves performance by 5.1–6.9 on settings with longer dialogue and more topic transitions, demonstrating the value of context-aware memory construction and query-conditioned visual expansion for long-horizon multimodal conversation.

1 Introduction

Vision-language models (VLMs) enable conversational agents to reason jointly over language and visual content (OpenAI, 2026; Anthropic, 2026). In long-horizon interactions, however, a single query may depend on information introduced many turns earlier. Re-processing the full history on every turn inflates latency, token cost, and input length, eventually exhausting the context budget. Worse, naive truncation silently discards visual evidence that later queries still need; when a user returns to a topic after a long digression, the precise visual details required to answer accurately are already gone. An effective multimodal memory must therefore preserve relevant visual evidence and retrieve it on demand, without repeatedly exposing the model to the entire interaction history.

Deciding what to preserve is difficult because images do not offer discrete semantic units like words. Meaningful components are implicit in continuous pixels (Ma et al., 2023), so one cannot simply extract and index the "important parts" without first determining what matters. Existing approaches fall into two camps, both inadequate. Text-only memory systems translate images into language, which inevitably drops visual attributes—texture, shape, color, spatial relations—that resist precise verbal description (Guo et al., 2026; Flickner et al., 1995; Todd & Petrov, 2022; Viethen et al., 2017). Multimodal memories retain the raw images, but typically index each one with a single global embedding that collapses fine-grained regions into an undifferentiated vector and fails to represent small areas that become relevant later (Kim et al., 2026; Li et al., 2026c). To capture the necessary granularity, one must decompose images into local regions as independently retrievable units, while preserving the full image for global context.

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.

Decomposition alone, however, only identifies *where* evidence is located, not *which* regions deserve to be remembered. An image may contain many visually coherent regions, but their value as conversational memory varies with the dialogue context. As shown in fig. 1, once prior utterances situate the image as a beverage advertisement, the multi-pack of Coca-Cola bottles becomes a salient entity worth preserving as an independent retrieval unit; absent that context, it is merely one object among many. The dialogue therefore does not merely accompany the image—it acts as a filter that determines which visual entities should be extracted, segmented, and indexed. Memory construction must leverage the context available at each image-bearing turn to select salient regions, anchor them to their parent turn, and store them as subimages with descriptive text so that both fine-grained evidence and its surrounding dialogue remain accessible at generation time.

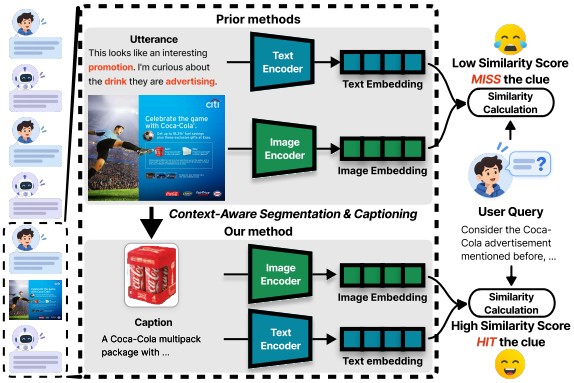


Figure 1: Context-aware memory construction in CoReMem. The dialogue situates the image as a beverage advertisement, so the method prioritizes the multi-pack of Coca-Cola bottles as a salient object and indexes its cropped image together with a text description alongside the original image. This context-selected subimage matches later queries more reliably than whole-image text–image matching that ignores the relevant object.

Retrieval raises a complementary challenge. A text query often cannot describe the visual evidence it needs: a terse "how much is it?" may depend on the exact appearance of an object introduced many turns ago, an appearance that no textual paraphrase can faithfully recover (Wei et al., 2024). Moreover, the evidence for a single answer is frequently distributed across multiple turns, and fixed top-k retrieval—scoring all candidates in one pass—cannot discover evidence that only becomes identifiable after intermediate turns have been retrieved (Qi et al., 2019; Jeong et al., 2024). Consequently, the retrieval process must be able to assess whether the evidence gathered so far is sufficient, and to search again when it is not.

We propose Context-Aware Memory Construction and Query-Aware Retrieval (CoReMem), built on the principle that both construction and retrieval should use high-level semantic signals—context and query intent, respectively—to determine what visual information is worth keeping and when to keep looking. During construction, the VLM uses the available dialogue context to identify salient regions; SAM3 (Carion et al., 2026) segments them into independent subimages with accompanying text. Each subimage remains linked to its parent turn, preserving the original image and surrounding dialogue for answer generation. During retrieval, CoReMem first fetches an initial set of turns via hybrid text–image matching, then prompts the VLM to evaluate whether the accumulated evidence suffices. If the evidence is insufficient, the system uses the images from the retrieved turns as visual queries to search the stored subimages, expanding context through iterative image-to-image retrieval. This design accesses fine-grained visual evidence beyond the reach of the initial text query, while allowing the retrieval depth to adapt to the evidence each specific question actually requires.

We evaluate CoReMem on MemEye (Guo et al., 2026), an image-centric benchmark for long-horizon multimodal conversation, using Qwen3.6-27B (Qwen Team, 2026a) and Qwen3.6-35B-A3B (Qwen Team, 2026b). With Qwen3.6-27B, CoReMem improves the overall MemEye judge score by 8.6 points over the strongest baseline and by 5.1–6.9 points on the concatenated tasks, which contain longer dialogues and more topic transitions. These results show that CoReMem retains an advantage over the evaluated baselines as histories become longer and more heterogeneous.

Our contributions are: (1) We treat conversational context as a selection signal for memory construction, indexing salient visual regions as independently retrievable, text-accompanied subimages rather than relying solely on global embeddings. (2) We introduce iterative, evidence-conditioned retrieval that assesses sufficiency and constructs visual queries from already-retrieved turns to search stored subimages, overcoming the limitations of fixed top-k and text-only retrieval. (3) We instantiate and evaluate this design on MemEye and its concatenated variants across model scales, demonstrating consistent gains that remain robust as dialogues grow longer and include more topic transitions.

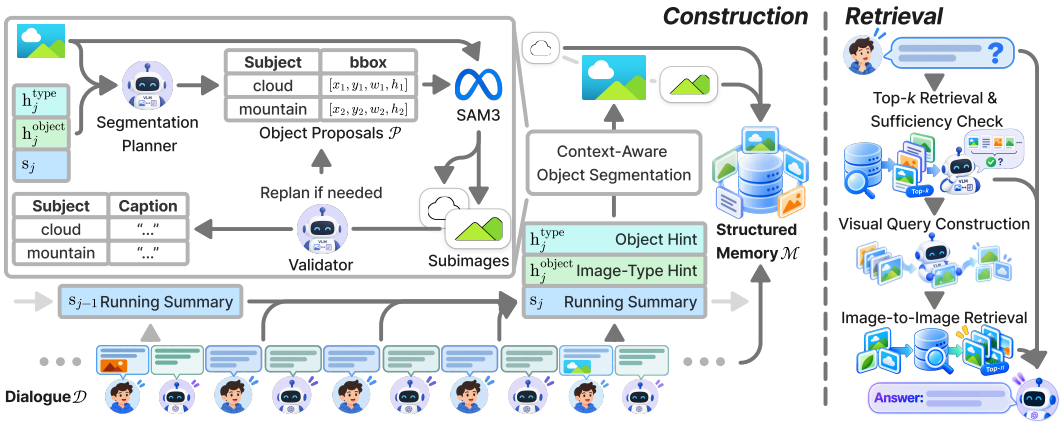


Figure 2: Overview of CoReMem. During memory construction, a VLM uses a running dialogue summary and image-specific hints to propose salient regions; SAM3 segments the proposal and a VLM validates the resulting mask. The accepted regions become subimage documents linked to their parent turns. During retrieval, hybrid text–image ranking selects an initial set of parent turns, a VLM checks whether the context is sufficient, and, when needed, visual queries constructed from retrieved images guide image-to-image retrieval over unseen turns to expand the context.

2 Method

We represent a dialogue as a sequence of turns $\mathcal{D} = (\mathrm{d}_1, \mathrm{d}_2, \ldots)$, where $\mathrm{d}_i = (\mathrm{u}_i, \mathrm{r}_i, \boldsymbol{I}_i)$ contains a user utterance u_i, an agent response r_i, and an optional user-provided image $\boldsymbol{I}_i$. We set $\boldsymbol{I}_i = \emptyset$ when no image is present and assume at most one image per turn. Figure 2 summarizes the method, and appendix C provides the system prompts used during memory construction and retrieval.

2.1 Memory Construction

Running dialogue summary and segmentation hints. To condition segmentation on conversational context, we process the dialogue preceding and including each image-bearing turn. Let $\mathbb{I}_{\text{image}} = \{i \mid \boldsymbol{I}_i \neq \emptyset\}$, let $m = |\mathbb{I}_{\text{image}}|$, and denote its ordered elements by $i_1 < i_2 < \cdots < i_m$. Setting $i_0 = 0$, we define the j-th local context block as the ordered sequence $\mathcal{D}_j^{\text{local}} := (\mathrm{d}_{i_{j-1}+1}, \ldots, \mathrm{d}_{i_j})$ for $j \in \{1, \ldots, m\}$. Across blocks, we maintain a running summary s_j of stable dialogue context. Given its previous value and the current block, the VLM updates this summary while producing an object hint $\mathrm{h}_j^{\text{object}}$ identifying visually grounded subjects and image-type hint $\mathrm{h}_j^{\text{type}}$, *e.g.,* `screenshot` or `landscape`, guiding type-appropriate region selection:

$$\mathrm{VLM}\big(\mathcal{D}_j^{\text{local}}, \mathrm{Prompt}_{\text{object-hint}}, \mathrm{s}_{j-1}\big) \to \big(\mathrm{s}_j, \mathrm{h}_j^{\text{object}}, \mathrm{h}_j^{\text{type}}\big), \tag{1}$$

where we set $\mathrm{s}_0 = \emptyset$. The segmentation planner then uses this context bundle—the updated summary and both hints—to process the image in d_{i_j}.

Context-aware object segmentation. For each image-bearing turn i_j, we set $\boldsymbol{I} = \boldsymbol{I}_{i_j}$. Conditioned on $\boldsymbol{I}$, the running summary, and the two hints, the VLM generates a set of object proposals $\mathcal{P}$. Each proposal $\mathrm{p} \in \mathcal{P}$ comprises a bounding box b and a label l:

$$\mathrm{VLM}\big(\boldsymbol{I}, \mathrm{s}_j, \mathrm{h}_j^{\text{object}}, \mathrm{h}_j^{\text{type}}, \mathrm{Prompt}_{\text{segmentation}}\big) \to \mathcal{P}. \tag{2}$$

For each proposal $\mathrm{p} = (\mathrm{b}, \mathrm{l})$, SAM3 receives the original image $\boldsymbol{I}$, bounding box b, and label l and produces an object mask $\boldsymbol{M}$. A VLM validates whether the mask isolates a coherent and potentially useful visual subject. For an accepted mask, the VLM generates a caption c:

$$\mathrm{VLM}\big(\boldsymbol{I}, \boldsymbol{M}, \mathrm{Prompt}_{\text{description}}\big) \to \mathrm{c}. \tag{3}$$

For a rejected mask, the VLM conditions on the image, current proposal, mask, and validator feedback f to overwrite the proposal or return a discard decision:

$$\mathrm{VLM}\big(\boldsymbol{I}, \mathrm{p}, \boldsymbol{M}, \mathrm{f}, \mathrm{Prompt}_{\text{adjustment}}\big) \to \mathrm{p} \quad \text{or} \quad \texttt{discard}. \tag{4}$$

We repeat this process until acceptance, discard, or the revision limit.

For each accepted mask, we crop its tight bounding rectangle from the original image, retain pixels inside the mask, and replace the remaining pixels with white, yielding subimage $\boldsymbol{I}_{\mathrm{sub}}$.

Memory embedding construction. For each turn $\mathrm{d} = (\mathrm{u}, \mathrm{r}, \boldsymbol{I})$, we embed the concatenation of user utterance u and agent response r in the text space and, when present, embed user-provided image $\boldsymbol{I}$ in the multimodal space. For each subimage, we embed its caption and pixels in the corresponding spaces. To support subsequent image-to-image retrieval, we additionally precompute and store DINOv3 (Siméoni et al., 2026) embeddings for all original images and subimages.

The resulting structured memory $\mathcal{M}$ contains two document types. A *turn document* stores a turn and its optional original image; a *subimage document* stores a segmented image, its caption, and its parent-turn index. Each turn contributes one turn document and zero or more subimage documents. Selecting a subimage document adds its parent turn to the response context while retaining the subimage document for further image-to-image retrieval.

2.2 Memory retrieval

The complete retrieval procedure is summarized in algorithm 1. A case study about a retrieval trace is provided in appendix A.

Initial hybrid retrieval. Given a text-only query q, we rank each document $x \in \mathcal{M}$ by averaging its cosine similarities to the query in the text and multimodal embedding spaces:

$$\mathrm{score}(x, \mathrm{q}) \coloneqq \frac{\mathrm{sim}_{\mathrm{text}}(x, \mathrm{q}) + \mathrm{sim}_{\mathrm{vis}}(x, \mathrm{q})}{2}. \tag{5}$$

Here, $\mathrm{sim}_{\mathrm{text}}$ and $\mathrm{sim}_{\mathrm{vis}}$ denote the text-space and multimodal-space similarities, respectively. For a document without an image, we set $\mathrm{sim}_{\mathrm{vis}}$ to zero.

The longest prefix of the ranked document list spanning at most k distinct parent turns defines $\mathcal{R}$. Here, $\mathrm{turns}(\cdot)$ maps a document set to its distinct parent turns. Initializing the accumulated document set as $\mathcal{S} = \mathcal{R}$ gives $\mathcal{C} = \mathrm{turns}(\mathcal{S})$, the accumulated turn-level response context.

VLM-based sufficiency check. After initial hybrid retrieval and after each subsequent image-to-image round, a VLM checks whether the accumulated context $\mathcal{C}$ is sufficient for query q. If it is, response generation begins; otherwise, retrieval continues for at most T rounds.

Visual query construction. When the context is insufficient, CoReMem constructs visual queries from image-bearing turns in $\mathrm{turns}(\mathcal{R})$. For each such turn, SoM (Yang et al., 2023) is used to mark any retrieved subimage segments on the original image. Conditioned on query q, the turn's user and agent messages, and the marked image, the VLM selects segments whose identity, appearance, state, recurrence, or change could retrieve additional evidence needed for the answer. If the marked segments do not cover such evidence, it proposes a new coherent, recognizable region that can be segmented by SAM3. New proposals follow section 2.1 without the running summary, hints, or proposal refinement. The resulting regions form $\mathcal{Q}$.

Image-to-Image retrieval. Let $\mathcal{M}_{\mathrm{img}} \coloneqq \{x \in \mathcal{M} \mid x \text{ stores an image}\}$. At each round, we exclude every document whose parent turn is already in $\mathcal{C}$. We compare each $v \in \mathcal{Q}$ with the remaining image-bearing candidates in $\mathcal{M}_{\mathrm{img}}$, using cosine similarity between its DINOv3 embedding and the stored document embeddings. We fuse the rankings with reciprocal rank fusion (RRF) (Cormack et al., 2009); for $x \in \mathcal{M}_{\mathrm{img}}$,

$$\mathrm{RRF}(x; \mathcal{Q}) \coloneqq \sum_{v \in \mathcal{Q}} \frac{1}{\kappa_{\mathrm{RRF}} + \mathrm{rank}_v(x)}, \tag{6}$$

where $\mathrm{rank}_v(x)$ is one-based and $\kappa_{\mathrm{RRF}} > 0$. If $\mathcal{Q} = \emptyset$ or no candidate is available, retrieval stops. Otherwise, as in the initial retrieval step, we assign to $\mathcal{R}$ the longest prefix of the RRF-ranked candidate list spanning at most n distinct parent turns. We then update $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{R}$ and $\mathcal{C} \leftarrow \mathrm{turns}(\mathcal{S})$, after which the VLM-based sufficiency check is repeated until success or the T-round budget is exhausted.

Algorithm 1: Query-aware multimodal memory retrieval

Input: Text-only query q, memory $\mathcal{M}$, initial turn budget k, per-iteration parent-turn budget n, iteration budget T
Output: Response r

```
/* Initial hybrid retrieval                                          */
```
foreach *document* $x \in \mathcal{M}$ **do** `// sim_vis = 0 if x has no image`
  $\mathrm{score}(x, \mathrm{q}) \leftarrow \big(\mathrm{sim}_{\mathrm{text}}(x, \mathrm{q}) + \mathrm{sim}_{\mathrm{vis}}(x, \mathrm{q})\big)/2$
$\mathcal{R} \leftarrow$ longest score-ranked prefix covering at most k distinct parent turns
$\mathcal{S} \leftarrow \mathcal{R}, \quad \mathcal{C} \leftarrow \mathrm{turns}(\mathcal{S})$ `// accumulated documents and response context`
```
/* VLM-based sufficiency check                                       */
```
if $\mathrm{VLM}\big(\mathcal{C}, \mathrm{q}, \mathrm{Prompt}_{\mathrm{sufficiency}}\big)$ *deems* $\mathcal{C}$ *sufficient* **then return** $\mathrm{VLM}(\mathcal{C}, \mathrm{q})$
for $t \leftarrow 1$ **to** T **do**
  ```
  /* Visual query construction                                       */
  ```
  $\mathcal{Q} \leftarrow \emptyset$
  foreach *image-bearing turn* $\mathrm{d}_i \in \mathrm{turns}(\mathcal{R})$ **do**
    mark retrieved subimage segments on the parent image $\boldsymbol{I}_i$ using SoM
    conditioned on q, use the VLM to select marked segments or propose new regions, and add the resulting regions to $\mathcal{Q}$
  if $\mathcal{Q} = \emptyset$ **then break**
  ```
  /* Image-to-image retrieval                                        */
  ```
  compute RRF scores only for image-bearing candidates whose parent turns are not in $\mathcal{C}$ using all $v \in \mathcal{Q}$ (eq. (6))
  $\mathcal{R} \leftarrow$ longest RRF-ranked prefix spanning at most n distinct parent turns
  if $\mathcal{R} = \emptyset$ **then break**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{R}, \quad \mathcal{C} \leftarrow \mathrm{turns}(\mathcal{S})$
  if $\mathrm{VLM}\big(\mathcal{C}, \mathrm{q}, \mathrm{Prompt}_{\mathrm{sufficiency}}\big)$ *deems* $\mathcal{C}$ *sufficient* **then return** $\mathrm{VLM}(\mathcal{C}, \mathrm{q})$
return $\mathrm{VLM}(\mathcal{C}, \mathrm{q})$ `// budget reached or no visual query/candidate`

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We evaluate all methods on MemEye (Guo et al., 2026) under two response-generation models, Qwen3.6-27B (Qwen Team, 2026a) and Qwen3.6-35B-A3B (Qwen Team, 2026b), using LLM-as-a-Judge (Zheng et al., 2023). A blinded human-agreement study found 95.1% agreement between automatic and human judgments across 144 responses, supporting the reliability of this evaluator; details are provided in appendix E. We report both the overall score and scores for MemEye's X and Y categories. X_1–X_4 require increasingly fine-grained visual evidence, whereas Y_1–Y_3 require increasingly complex multi-hop reasoning over interdependent visual evidence (Guo et al., 2026). Further dataset details are provided in appendix D.

We compare CoReMem with text-only and multimodal memory baselines. The text-only group comprises A-Mem (Xu et al., 2025), MemoryOS (Kang et al., 2025), LightRAG (Guo et al., 2025), SimpleMem (Liu et al., 2026b), and caption-based text RAG (TextRAG) (Lewis et al., 2020; Karpukhin et al., 2020). To apply these methods to MemEye, we append a detailed caption generated by the response-generation VLM to each image-bearing dialogue turn, so memory construction and retrieval operate on textual representations. The multimodal group comprises Omni-SimpleMem (Liu et al., 2026a), MMA (Lu et al., 2026), M^{2}A (Feng et al., 2026), and multimodal RAG (MMRAG), which retain or directly encode visual evidence. MMRAG augments text retrieval with siglip2-base-patch16-384 (Tschannen et al., 2025) as its image encoder. All methods initially retrieve the top-10 dialogue turns. We use all-MiniLM-L6-v2 (Wang et al., 2020) as the text encoder and models from the SigLIP family (Zhai et al., 2023; Tschannen et al., 2025) as the multimodal encoders. For CoReMem, each image-to-image retrieval round returns a ranked prefix spanning at most three previously unseen parent turns ($n = 3$), and we set the RRF constant κ_{RRF} to 20. Unless otherwise specified, we cap both segmentation revisions and image-to-image retrieval iterations at 3. Further implementation details are provided in appendix F.

Table 1: LLM-as-a-Judge scores reported (0–100, ↑) on MemEye. X_1–X_4 assess increasingly fine-grained visual evidence; Y_1–Y_3 assess increasingly complex multihop reasoning. Methods are grouped by memory representation: text-only methods index textual representations, including generated image captions, whereas multimodal methods retain or directly encode visual evidence. Results are marked as best and second best within each model.

Model	Modality	Method	Overall	*X* Category				*Y* Category		
				X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
Qwen3.6-27B	Text-only	A-Mem	48.1	50.0	46.9	47.2	48.9	55.8	43.8	46.9
		MemoryOS	36.4	35.4	21.9	39.1	39.1	37.1	34.7	41.7
		LightRAG	42.6	66.7	35.4	51.4	27.1	56.7	37.2	30.2
		SimpleMem	31.9	46.9	28.1	26.1	34.2	28.7	32.3	38.5
		Text RAG	47.4	58.3	44.8	54.2	37.2	61.3	41.9	36.5
	Multimodal	MMA	46.2	53.1	41.7	58.1	32.7	62.9	40.9	27.1
		M^{2}A	32.2	40.6	20.8	35.6	29.7	30.0	34.0	30.2
		Omni-SimpleMem	45.6	65.6	43.8	51.4	32.7	56.7	42.9	29.2
		MMRAG	51.2	56.3	45.8	59.2	42.9	64.6	47.0	35.4
		Ours	59.8	60.4	63.5	67.6	50.0	68.3	58.6	43.8
Qwen3.6-35B-A3B	Text-only	A-Mem	39.2	49.0	31.3	43.7	33.8	45.8	34.7	41.7
		MemoryOS	31.3	28.1	17.7	34.9	33.5	35.0	29.6	29.2
		LightRAG	35.6	53.1	31.2	41.5	24.4	48.3	27.6	37.5
		SimpleMem	30.1	43.8	31.2	27.8	27.1	34.2	28.1	28.1
		Text RAG	47.9	56.3	37.5	54.2	41.7	61.3	40.6	44.8
	Multimodal	MMA	32.9	40.6	14.6	38.4	30.8	42.9	30.3	18.8
		M^{2}A	27.5	25.0	20.8	27.5	30.8	25.0	27.3	34.4
		Omni-SimpleMem	46.5	68.8	41.7	51.1	35.3	56.2	42.4	39.6
		MMRAG	48.6	57.3	49.0	56.3	36.8	65.4	42.9	30.2
		Ours	52.0	58.3	47.9	59.2	43.6	68.8	44.3	42.7

Table 2: LLM-as-a-Judge scores reported (0–100, ↑) on the Concat, Quad, and Full concatenated MemEye variants. Results are marked as best and second best within each variant and model.

Variant	Qwen3.6-27B									
	A-Mem	MemoryOS	LightRAG	SimpleMem	Omni-SimpleMem	MMA	M^{2}A	Text RAG	MMRAG	Ours
Concat	48.1	33.3	41.9	36.8	46.9	47.3	32.2	46.9	50.3	55.4
Quad	48.6	33.5	45.4	31.5	49.3	43.9	29.7	45.0	45.8	56.2
Full	49.2	29.9	41.1	28.6	47.7	43.8	30.9	45.0	46.2	55.8
Variant	Qwen3.6-35B-A3B									
	A-Mem	MemoryOS	LightRAG	SimpleMem	Omni-SimpleMem	MMA	M^{2}A	Text RAG	MMRAG	Ours
Concat	41.9	34.6	35.8	34.4	44.3	33.7	26.7	45.0	47.2	52.7
Quad	41.5	35.3	35.2	30.9	43.7	32.3	25.9	44.0	44.6	49.2
Full	40.3	30.4	35.8	30.7	44.7	29.6	23.9	43.3	44.9	49.2

3.2 Main results

Table 1 shows that CoReMem achieves the strongest overall score with both models, improving over MMRAG by 8.6 with Qwen3.6-27B and 3.4 with Qwen3.6-35B-A3B.[1] The category pattern points to complementary contributions from memory construction and retrieval. Context-aware subimages make localized evidence independently searchable, while the sufficiency check and query-aware image-to-image expansion retrieve additional evidence when the initial context is inadequate. Consistent with this design, CoReMem leads on the fine-grained X_3–X_4 categories for both models, with gains over the strongest alternative of $8.4/1.1$ and $2.9/1.9$, respectively, for Qwen3.6-27B and Qwen3.6-35B-A3B; it also leads on Y_1–Y_2 at both scales, with gains of $3.7/11.6$ and $3.4/1.4$.

[1] An MoE model with 3B activations per token.

On Y_3, the best text-only method remains ahead, but CoReMem narrows the multimodal gap: relative to the best text-only result, the deficit is 3.1 with Qwen3.6-27B and 2.1 with Qwen3.6-35B-A3B, compared with 11.5 and 5.2 for the best prior multimodal method. Thus, the combination of fine-grained memory units and adaptive visual retrieval improves both localized evidence access and multi-hop reasoning for long-horizon multimodal conversation.

3.3 Results on extended tasks

To evaluate retrieval over longer contexts, we concatenate MemEye's eight scenario dialogues into three variants: *Concat* uses disjoint pairs, *Quad* uses disjoint groups of four, and *Full* uses all eight dialogues. These variants increase both dialogue length and topic transitions.

Robustness to extended contexts. The results in table 2 show that longer histories do not eliminate the advantage of CoReMem: its margin over the strongest alternative remains 4.3–6.9 across all settings. On the Full variant, the margins are 6.6 with Qwen3.6-27B and 4.3 with Qwen3.6-35B-A3B. Performance is nearly unchanged from Concat to Full with the 27B model (55.4 versus 55.8); with the 35B model, the initial drop from 52.7 to 49.2 is not compounded from Quad to Full. The relative lead persists as context expands.

Performance–context trade-off. We vary retrieval budgets on the MemEye Full setting and compare CoReMem with MMRAG, TextRAG, and MMRAG augmented with Qwen3-VL-Reranker-8B (Li et al., 2026b), denoted MMReranker. MMReranker first retrieves the top-25 candidates with MMRAG and then reranks this fixed candidate pool; the other baselines vary top-k directly. For CoReMem, the six points use iteration budgets $T = 0, \ldots, 5$. As shown in fig. 3, CoReMem generally scores higher at comparable context sizes and achieves its best score with fewer turns.

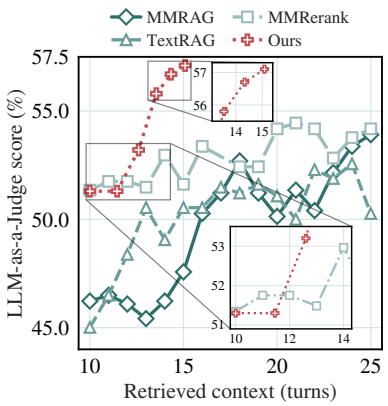


Figure 3: Performance on the MemEye Full setting as a function of the average number of retrieved turns. From left to right, the CoReMem points use iteration budgets $T = 0, \ldots, 5$; the baseline curves vary the top-k budget.

3.4 Cost analysis

We compare completion-token usage during memory construction for agentic memory methods on MemEye Full setting with Qwen3.6-27B. As shown in fig. 4, CoReMem attains the highest LLM-as-a-Judge score while consuming 5.8×10^5 completion tokens. This is less than half the usage of Omni-SimpleMem (1.4×10^6) and A-Mem (1.2×10^6), although it is $1.3\times$ higher than MemoryOS (4.4×10^5).

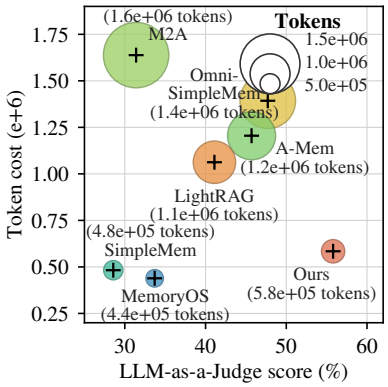


Figure 4: Performance–token cost comparison during memory construction on MemEye Full setting with Qwen3.6-27B.

4 Method analysis

4.1 Mechanism analysis

Figure 5 summarizes mechanism statistics collected with Qwen3.6-27B on MemEye. Panel (a) reports the fraction of retrieved documents that are subimage documents. The higher proportions for X_2–X_4 show that subimage documents more often rank within the retrieved set when queries require fine-grained visual evidence. Panel (b) reports the mean number of image-to-image retrieval rounds. Although the iteration budget is $T = 3$, the sufficiency check terminates retrieval early for many queries, yielding category-wise means of only 0.40–1.70 rounds. Panel (c) reports the mean number of regions dynamically proposed during retrieval. Higher averages for Y_2 and Y_3 suggest that the system proposes more regions when a question requires multiple evidence items.

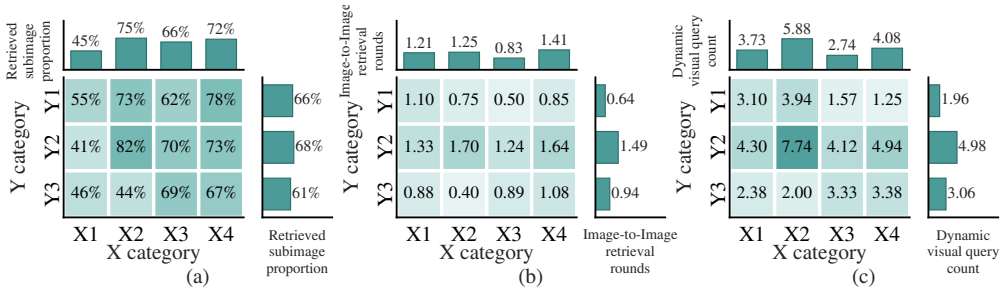


Figure 5: Mechanism statistics collected with Qwen3.6-27B on MemEye. From left to right, panels report the retrieved subimage-document proportion, the mean number of image-to-image retrieval rounds, and the mean number of regions dynamically proposed during retrieval.

Table 3: Component ablations of CoReMem and comparison with MMRAG on MemEye using Qwen3.6-27B. LLM-as-a-Judge scores are reported (0–100, ↑). Lower-right annotations report changes relative to MMRAG. Results are marked as best and second best within each column.

Configuration	Overall	X Category				Y Category		
		X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
MMRAG	51.2	56.3	45.8	59.2	42.9	64.6	47.0	35.4
Full CoReMem	59.8$_{\uparrow 8.6}$	60.4$_{\uparrow 4.1}$	63.5$_{\uparrow 17.7}$	67.6$_{\uparrow 8.4}$	50.0$_{\uparrow 7.1}$	68.3$_{\uparrow 3.7}$	58.6$_{\uparrow 11.6}$	43.8$_{\uparrow 8.4}$
Memory construction phase								
w/o Object segmentation	55.8$_{\uparrow 4.6}$	62.5$_{\uparrow 6.2}$	54.2$_{\uparrow 8.4}$	63.4$_{\uparrow 4.2}$	45.9$_{\uparrow 3.0}$	72.1$_{\uparrow 7.5}$	51.0$_{\uparrow 4.0}$	35.4$_{0.0}$
w/o Summary and hints	56.6$_{\uparrow 5.4}$	65.6$_{\uparrow 9.3}$	60.4$_{\uparrow 14.6}$	62.7$_{\uparrow 3.5}$	45.5$_{\uparrow 2.6}$	69.2$_{\uparrow 4.6}$	51.7$_{\uparrow 4.7}$	45.8$_{\uparrow 10.4}$
Memory retrieval phase								
w/o Iterative retrieval	54.9$_{\uparrow 3.7}$	54.2$_{\downarrow 2.1}$	53.1$_{\uparrow 7.3}$	65.1$_{\uparrow 5.9}$	44.7$_{\uparrow 1.8}$	64.2$_{\downarrow 0.4}$	51.2$_{\uparrow 4.2}$	46.9$_{\uparrow 11.5}$
w/o VLM-based sufficiency check	59.2$_{\uparrow 8.0}$	58.3$_{\uparrow 2.0}$	54.2$_{\uparrow 8.4}$	66.5$_{\uparrow 7.3}$	53.4$_{\uparrow 10.5}$	71.3$_{\uparrow 6.7}$	55.4$_{\uparrow 8.4}$	44.8$_{\uparrow 9.4}$
w/o Dynamic query construction	52.4$_{\uparrow 1.2}$	52.1$_{\downarrow 4.2}$	51.0$_{\uparrow 5.2}$	63.0$_{\uparrow 3.8}$	41.7$_{\downarrow 1.2}$	62.9$_{\downarrow 1.7}$	49.3$_{\uparrow 2.3}$	39.6$_{\uparrow 4.2}$

4.2 Ablation Studies

Memory construction phase. Removing object segmentation lowers the overall score by 4.0, with the largest drops on X_2 (9.3) and Y_3 (8.4). Removing the running summary and both hints lowers the overall score by 3.2, mainly on X_2–X_4 and Y_2, but improves X_1 and Y_3. These results show that the construction components improve aggregate performance, but not uniformly across categories.

Memory retrieval phase. Removing the iterative retrieval process lowers the overall score by 4.9, whereas removing dynamic query construction produces the largest overall loss (7.4) and reduces every category score. Removing the VLM-based sufficiency check changes the overall score by only 0.6. Together with the early termination observed in fig. 5(b), this result indicates that the VLM-based sufficiency check primarily controls the number of retrieval rounds rather than uniformly improving accuracy.

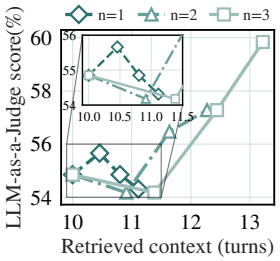


Figure 6: Sensitivity to the per-iteration parent-turn budget n and iteration budget T on MemEye with Qwen3.6-27B.

4.3 Hyperparameter Analysis

We evaluate the per-iteration parent-turn budget $n \in \{1, 2, 3\}$ and iteration budget $T \in \{0, \ldots, 3\}$ on the MemEye Full setting with Qwen3.6-27B (fig. 6). Increasing iteration budget T yields little consistent improvement when per-iteration parent-turn budget $n = 1$, whereas larger per-iteration parent-turn budgets n benefit more from additional iteration budget T. This trend suggests that iterative expansion is most effective when each round supplies a sufficiently broad candidate set.

5 Related work

Conversational memory. Early conversational-memory research predominantly focused on text-only interactions, representing dialogue histories and stored memories in text form. This line of work established benchmarks for long-term interaction and explored mechanisms for organizing information across turns (Wu et al., 2025; Tan et al., 2025; Hu et al., 2026; Tavakoli et al., 2025). MemoryBank (Zhong et al., 2024) maintains event summaries and user profiles, MemoryOS (Kang et al., 2025) organizes memories hierarchically, and A-Mem (Xu et al., 2025) adopts Zettelkasten-inspired, context-aware memory management. Together, these systems characterize key design choices in memory construction and retrieval, while leaving the preservation and retrieval of visual evidence outside their primary scope.

As conversational agents increasingly handle images alongside language, recent work has extended conversational memory to multimodal interactions (Bei et al., 2026; Guo et al., 2026). M^2A (Feng et al., 2026) separates raw messages from semantic memories, MMA (Lu et al., 2026) models the reliability of multimodal evidence, and Omni-SimpleMem (Liu et al., 2026a) combines dense, sparse, and graph-based search, linking caption-based retrieval back to the original image. In contrast, CoReMem uses dialogue context to construct independently searchable subimage memories while retaining their parent turns for answer generation. It further supplements initial text–image retrieval with sufficiency-controlled, query-conditioned visual retrieval.

Fine-grained visual representation and retrieval. Several methods represent or retrieve visual evidence at finer granularity. REVIVE demonstrates the value of regional visual representations for knowledge-based VQA (Lin et al., 2022), while FES-RAG selects sentence- and region-level fragments from retrieved documents (Wang et al., 2026). RegionRAG and PixSearch further investigate region-level and pixel-grounded retrieval for visual document understanding and knowledge-intensive multimodal reasoning (Li et al., 2026c; Kim et al., 2026). CoReMem differs in both setting and construction: it uses conversational context to identify salient regions when an image enters memory, indexes the resulting subimages independently, and preserves links to their parent turns.

Adaptive and iterative retrieval. Adaptive retrieval methods replace a fixed retrieval budget with decisions conditioned on the evolving generation context. IRCoT (Trivedi et al., 2023) interleaves retrieval with chain-of-thought reasoning for multistep questions. FLARE (Jiang et al., 2023) anticipates upcoming sentences and retrieves when generation confidence is low, whereas Self-RAG (Asai et al., 2024) learns to retrieve and critique generations through reflection tokens. S2G-RAG (Li et al., 2026a) explicitly assesses evidence sufficiency, identifies information gaps, and formulates targeted follow-up queries. These methods primarily retrieve textual knowledge. CoReMem instead applies sufficiency-controlled retrieval to multimodal conversational memory and expands the retrieved context through query-conditioned image-to-image search.

6 Conclusion

We presented CoReMem, a multimodal conversational memory framework that combines context-aware memory construction with query-aware retrieval. It indexes dialogue-relevant image regions as subimage documents linked to their parent turns, then uses sufficiency-controlled visual expansion to retrieve additional evidence when the initial text–image context is inadequate.

On MemEye, CoReMem improves over the strongest evaluated baseline by 8.6 points with Qwen3.6-27B and by 3.4 points with Qwen3.6-35B-A3B, with gains on fine-grained visual evidence, multi-hop reasoning, and longer concatenated dialogues with more topic transitions. Ablations confirm the value of region-level memory construction and adaptive visual retrieval, while the VLM-based sufficiency check helps avoid unnecessary iterations.

Our current study is limited to text-and-image conversations. As AI systems evolve, conversational memory will increasingly need to represent and retrieve evidence across additional modalities, including speech, documents, and video. Progress in this direction is currently constrained by the limited availability of rich, long-horizon multimodal datasets; extending memory construction and retrieval to these modalities is therefore an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on the research methodology and experiments, to implement the proposed methods, to support qualitative analysis of experimental outcomes, to clean and reformat data sets, and to interpret results. We have not used generative AI tools to propose or refine hypotheses, or to assist with translation. Generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving those claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create and modify scientific figures, draft parts of the manuscript, summarize and analyse existing literature, identify research topics and gaps, brainstorm ideas, search for information, identify relevant literature, edit the manuscript for readability, suggest a paper structure, and propose a title and keywords. The authors reviewed all AI-assisted material: implementations and analyses were verified against the reported procedures and results; suggested literature was confirmed against the cited sources; and all figures and manuscript text were revised and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies memory construction and retrieval for multimodal conversational agents. We evaluate exclusively on the official MemEye release (Guo et al., 2026) and on concatenated variants of its existing dialogues; we do not collect new user data, regenerate benchmark instances, or release a new dataset.

The images used in our experiments are documented in appendix D. CrossScene Memory and Health Care use AI-generated scenes and dashboards rather than photographs of real interiors or clinical records; Card Playlog uses HTML-rendered game screenshots; and Social Chat composites twelve synthetic face avatars from an MIT-licensed StyleGAN dataset (javi22, 2025; Karras et al., 2020) into rendered chat interfaces. These avatars do not depict any real person. The remaining tasks use archival advertisements, public-domain comics, stock interior images, and openly released dashcam frames. We do not redistribute those assets.

The only human annotation in this work is a blinded agreement study in which one annotator judged the correctness of model responses against MemEye reference answers (appendix E). No interaction with human users of the system was conducted, and no personal data were collected from annotators beyond these correctness labels.

On the MemEye evaluation data, person-centric images are synthetic avatars and health-related screenshots are AI-generated, so the reported experiments do not use photographs of identifiable people or real patient records. CoReMem nevertheless stores original images and cropped subimages so that visual evidence can be retrieved later. If applied to user-provided photos, this design can retain sensitive content such as faces, home interiors, health-related screenshots, or location cues. The method is intended for research on conversational memory; deploying it on personal or clinical data would require consent, access control, retention limits, and other data-governance measures that are outside the scope of this paper. Because the system can index people and objects across long histories, it could be misused for persistent visual tracking; we do not support such use. The visual encoders and VLMs used here may also inherit biases from their training data, and MemEye's constructed life scenarios are not a representative sample of users or populations.

REPRODUCIBILITY STATEMENT

Anonymous source code for CoReMem and the experimental pipeline is included in the supplementary materials. The procedure, prompts, implementation choices, data, and evaluation protocol used in this work are specified in the paper and appendix. Section 2 defines context-aware memory construction and query-aware retrieval, algorithm 1 gives the complete retrieval procedure, and appendix B lists the notation. The VLM system prompts for construction and retrieval are provided in appendix C. Model checkpoints, encoders, decoding settings, CoReMem hyperparameters, and the implementation choices for every baseline are given in appendix F; the pinned vLLM images and serving flags are recorded in appendix G. All experiments use the official MemEye release (Guo

et al., 2026). Appendix D describes the benchmark, the Concat, Quad, and Full history constructions, and the open-ended LLM-as-a-Judge protocol, and appendix E reports the blinded human-agreement study used to check that evaluator. Main and extended-task results, the retrieval-budget sweep, and the construction-cost comparison are reported in section 3; component ablations and the sensitivity to the per-iteration parent-turn budget and the iteration budget are reported in table 3 and fig. 6.

REFERENCES

Anthropic. Introducing Claude Fable 5.1 and Claude Mythos 5.1. https://www.anthropic.com/claude-fable-and-mythos-5-1, September 2026. Accessed: 2026-09-05.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik F. Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 40750–40784. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1892. URL https://doi.org/10.18653/v1/2026.acl-long.1892.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel (eds.), *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pp. 758–759. ACM, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, and Wentao Zhang. M2A: multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *CoRR*, abs/2602.07624, 2026. doi: 10.48550/ARXIV.2602.07624. URL https://doi.org/10.48550/arXiv.2602.07624.

Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, September 1995. ISSN 0018-9162. doi: 10.1109/2.410146. URL https://doi.org/10.1109/2.410146.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, and Ruixiang Tang. Memeye: A visual-centric evaluation framework for multimodal agent memory. *CoRR*, abs/2605.15128, 2026. doi: 10.48550/ARXIV.2605.15128. URL https://doi.org/10.48550/arXiv.2605.15128.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 10746–10761. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.568. URL https://doi.org/10.18653/v1/2025.findings-emnlp.568.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1100–1110. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.123. URL https://doi.org/10.1109/CVPR.2017.123.

javi22. this-person-does-not-exist-10k. Hugging Face Datasets, 2025. URL https://huggingface.co/datasets/javi22/this-person-does-not-exist-10k. MIT License. Accessed: 2026-09-22.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.389. URL https://doi.org/10.18653/v1/2024.naacl-long.389.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.495. URL https://doi.org/10.18653/v1/2023.emnlp-main.495.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 25961–25970. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1318. URL https://doi.org/10.18653/v1/2025.emnlp-main.1318.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550. URL https://doi.org/10.18653/v1/2020.emnlp-main.550.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 8107–8116. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00813. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html.

Jeonghwan Kim, Renjie Tao, Sanat Sharma, Jiaqi Wang, Kai Sun, Zhaojiang Lin, Seungwhan Moon, Lambert Mathias, Anuj Kumar, Heng Ji, and Xin Luna Dong. Pixel-grounded retrieval for knowledgeable large multimodal models. *CoRR*, abs/2601.19060, 2026. doi: 10.48550/ARXIV.2601.19060. URL https://doi.org/10.48550/arXiv.2601.19060.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles,*

SOSP 2023, Koblenz, Germany, October 23-26, 2023, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Danrui Li, Sen Zhang, Samuel S. Sohn, Kaidong Hu, Muhammad Usman, and Mubbasir Kapadia. Cardiverse: Harnessing llms for novel card game prototyping. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 29735–29762. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1511. URL https://doi.org/10.18653/v1/2025.emnlp-main.1511.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: structured sufficiency and gap judging for iterative retrieval-augmented QA. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 25846–25862. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.ACL-LONG.1185. URL https://doi.org/10.18653/v1/2026.acl-long.1185.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *CoRR*, abs/2601.04720, 2026b. doi: 10.48550/ARXIV.2601.04720. URL https://doi.org/10.48550/arXiv.2601.04720.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. Regionrag: Region-level retrieval-augmented generation for visual document understanding. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 6662–6670. AAAI Press, 2026c. doi: 10.1609/AAAI.V40I8.37597. URL https://doi.org/10.1609/aaai.v40i8.37597.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. REVIVE: regional visual representation matters in knowledge-based visual question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/44956951349095f74492a5471128a7e0-Abstract-Conference.html.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, Mingyu Ding, and Huaxiu Yao. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory. *CoRR*, abs/2604.01007, 2026a. doi: 10.48550/ARXIV.2604.01007. URL https://doi.org/10.48550/arXiv.2604.01007.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for LLM agents. *CoRR*, abs/2601.02553, 2026b. doi: 10.48550/ARXIV.2601.02553. URL https://doi.org/10.48550/arXiv.2601.02553.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. MMA: multimodal memory agent. *CoRR*, abs/2602.16493, 2026. doi: 10.48550/ARXIV.2602.16493. URL https://doi.org/10.48550/arXiv.2602.16493.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ CREPE: can vision-language foundation models reason compositionally? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10910–10921. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01050. URL https://doi.org/10.1109/CVPR52729.2023.01050.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, September 2026. Accessed: 2026-09-05.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. Answering complex open-domain questions through iterative query generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2590–2602. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1261. URL https://doi.org/10.18653/v1/D19-1261.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b. Accessed: 2026-09-09.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b. Accessed: 2026-09-05.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 19336–19352. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.989. URL https://doi.org/10.18653/v1/2025.findings-acl.989.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *CoRR*, abs/2510.27246, 2025. doi: 10.48550/ARXIV.2510.27246. URL https://doi.org/10.48550/arXiv.2510.27246.

James T. Todd and Alexander A. Petrov. The many facets of shape. *Journal of Vision*, 22(1):1–1, 01 2022. ISSN 1534-7362. doi: 10.1167/jov.22.1.1. URL https://doi.org/10.1167/jov.22.1.1.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.557. URL https://doi.org/10.18653/v1/2023.acl-long.557.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/ARXIV.2502.14786. URL https://doi.org/10.48550/arXiv.2502.14786.

turing-motors. Japan-open-driving-dataset-sample. Hugging Face Datasets, 2026. URL https://huggingface.co/datasets/turing-motors/Japan-Open-Driving-Dataset-Sample. Accessed: 2026-09-25.

Jette Viethen, Thomas van Vessem, Martijn Goudbeek, and Emiel Krahmer. Color in reference production: the role of color similarity and color codability. *Cognitive science*, 41:1493–1514, 2017.

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. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Xihang Wang, Zihan Wang, Chengkai Huang, Cao Liu, Ke Zeng, Quan Z. Sheng, and Lina Yao. Purifying multimodal retrieval: Fragment-level evidence selection for RAG. In Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (eds.), *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, pp. 1880–1891. ACM, 2026. doi: 10.1145/3805712.3809692. URL https://doi.org/10.1145/3805712.3809692.

Zhaohui Wei, Lizi Liao, Xiaoyu Du, and Xinguang Xiang. Balancing visual context understanding in dialogue for image retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 7929–7942. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.465. URL https://doi.org/10.18653/v1/2024.findings-emnlp.465.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *CoRR*, abs/2310.11441, 2023. doi: 10.48550/ARXIV.2310.11441. URL https://doi.org/10.48550/arXiv.2310.11441.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025 - Workshops, Honolulu, HI, USA, October 19-20, 2025*, pp. 1871–1881. IEEE, 2025. doi: 10.1109/ICCVW69036.2025.00197. URL https://doi.org/10.1109/ICCVW69036.2025.00197.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion

Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 19724–19731. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

CONTENTS OF APPENDIX

A Case study **18**

B Notation **19**

C System prompts **20**

C.1 Memory construction 20

C.2 Memory retrieval 24

C.3 Answer generation 25

C.4 LLM-as-a-Judge evaluation 27

D The MemEye benchmark **29**

D.1 Overview 29

D.2 Two-dimensional taxonomy 29

D.3 Dataset statistics 29

D.4 Benchmark construction and evaluation 31

D.5 Representative examples 31

D.6 Detailed task descriptions 31

D.7 Image sources and privacy 33

E Human validation of automatic evaluation **33**

F Implementation details **33**

F.1 Models and encoders 33

F.2 CoReMem hyperparameters 34

F.3 Baseline implementations 34

G vLLM serving configuration **35**

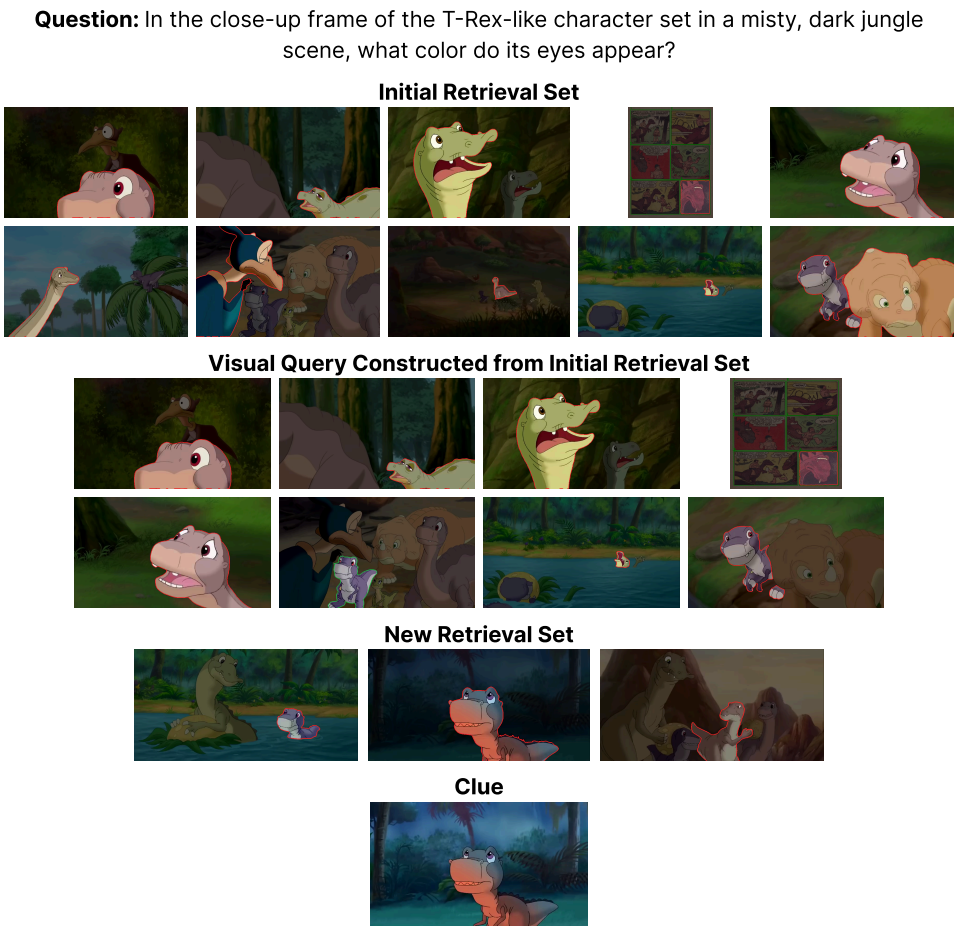


Figure 7: A retrieval trace for a Cartoon Entertainment query. The top section, *Initial Retrieval Set*, shows the initially retrieved parent images with red contours marking the selected subimages. The middle section, *Visual Query Constructed from Initial Retrieval Set*, uses these red-contoured subimages as visual queries and includes one additional region proposed dynamically, marked in green. The bottom section, *New Retrieval Set*, shows three images retrieved by image-to-image search; red contours mark the subimages matched by the visual queries. The final *Clue* panel shows the evidence needed to answer the query.

A Case study

Figure 7 illustrates one retrieval trace from the Cartoon Entertainment scene of MemEye (Guo et al., 2026). The query asks: *In the close-up frame of the T-Rex-like character set in a misty, dark jungle scene, what color do its eyes appear?* The top section shows the initial retrieval set. Red contours identify the subimages selected from the retrieved parent images; multiple contours on one image indicate that more than one subimage from the same turn was retrieved. None of these subimages contains the clue needed to answer the query, so the VLM judges the initial context insufficient.

The middle section shows the visual queries constructed from the initial retrieval set: the red-contoured subimages are reused as visual queries, and one additional region is proposed dynamically, marked in green. Image-to-image retrieval then produces the three images in the *New Retrieval Set*; the red contours in these images identify the subimages matched by the visual queries. The *Clue* panel isolates the retrieved image containing the missing visual evidence, after which the accumulated context is judged sufficient and passed to the response generator. This example illustrates how query-conditioned visual expansion can recover evidence that is absent from the initial text–image retrieval.

B NOTATION

This appendix lists the symbols used for memory construction and retrieval. Indices i and j denote a dialogue turn and a local context block, respectively.

Table 4: Notation for memory construction and retrieval.

Symbol	Description
Dialogue and memory	
$\mathcal{D}$	A dialogue, represented as an ordered sequence of turns $(\mathrm{d}_1, \mathrm{d}_2, \ldots)$
d_i	The i-th turn $\mathrm{d}_i = (\mathrm{u}_i, \mathrm{r}_i, \boldsymbol{I}_i)$
$\mathrm{u}_i, \mathrm{r}_i$	The user utterance and the agent response in turn i
$\boldsymbol{I}_i$	An optional user-provided image in turn i; $\boldsymbol{I}_i = \emptyset$ if absent. At most one image per turn. $\boldsymbol{I} = \boldsymbol{I}_{i_j}$ denotes the current image
$\mathcal{M}$	The constructed memory: each turn contributes one turn document and zero or more subimage documents, each linked to exactly one parent turn
q	A text-only user query
Memory construction	
$\mathbb{I}_{\mathrm{image}}$	Indices of image-bearing turns, $\mathbb{I}_{\mathrm{image}} = \{i \mid \boldsymbol{I}_i \neq \emptyset\}$
m	The number of image-bearing turns, $m = \lvert\mathbb{I}_{\mathrm{image}}\rvert$
i_j	The index of the j-th image-bearing turn, with $i_1 < \cdots < i_m$ and $i_0 = 0$
$\mathcal{D}_j^{\mathrm{local}}$	The local context block ending at the j-th image-bearing turn, $\mathcal{D}_j^{\mathrm{local}} = (\mathrm{d}_{i_{j-1}+1}, \ldots, \mathrm{d}_{i_j})$
s_j	The running dialogue summary after processing block j; $\mathrm{s}_0 = \emptyset$
$\mathrm{h}_j^{\mathrm{object}}$	The object hint for local context block j
$\mathrm{h}_j^{\mathrm{type}}$	The image-type hint for local context block j
$\mathcal{P}$	The set of main-object proposals for the current image
$\mathrm{p} = (\mathrm{b}, \mathrm{l})$	A main-object proposal, comprising a bounding box b and an object label l
$\boldsymbol{M}$	An object mask produced by SAM3 (Carion et al., 2026)
c	A caption generated for an accepted mask
f	Validator feedback used to revise a rejected proposal
$\boldsymbol{I}_{\mathrm{sub}}$	An object-level subimage: the tight crop of an accepted mask, with pixels outside the mask replaced by white
Prompt_*	VLM system prompts object-hint, segmentation, description, adjustment, and sufficiency; see appendix C
Memory retrieval	
x	A memory document in $\mathcal{M}$
$\mathrm{sim}_{\mathrm{text}}$	Cosine similarity in the text embedding space
$\mathrm{sim}_{\mathrm{vis}}$	Cosine similarity in the multimodal embedding space; 0 if the document has no image
$\mathrm{score}(x, \mathrm{q})$	Hybrid retrieval score, the average of $\mathrm{sim}_{\mathrm{text}}(x, \mathrm{q})$ and $\mathrm{sim}_{\mathrm{vis}}(x, \mathrm{q})$
$\mathcal{R}$	The documents retrieved in the current iteration
$\mathcal{S}$	The documents accumulated across retrieval iterations
$\mathrm{turns}(\cdot)$	The parent turns associated with a document set
$\mathcal{C}$	The accumulated turn-level response context, $\mathcal{C} = \mathrm{turns}(\mathcal{S})$
$\mathcal{Q}$	The set of visual queries used for image-to-image retrieval
v	A visual query in $\mathcal{Q}$
$\mathcal{M}_{\mathrm{img}}$	The image-bearing documents in $\mathcal{M}$
$\mathrm{rank}_v(x)$	The one-based rank of document x under visual query v
$\mathrm{RRF}(x; \mathcal{Q})$	Reciprocal rank fusion score of x over the visual queries in $\mathcal{Q}$ (Cormack et al., 2009)
κ_{RRF}	The rank constant used in reciprocal rank fusion
k, n, T	Initial turn budget, per-iteration parent-turn budget, and maximum number of image-to-image retrieval rounds
Evaluation	
X_1–X_4	MemEye (Guo et al., 2026) visual-evidence categories, from scene-level to pixel-level
Y_1–Y_3	MemEye reasoning-depth categories, from atomic retrieval to evolutionary synthesis

C SYSTEM PROMPTS

This appendix provides the prompts used for memory construction, retrieval, final answer generation, and automatic evaluation. Unless otherwise noted, a listing is a *system* prompt that specifies the model's role, decision criteria, and output schema. For method-internal calls, the corresponding *user* message is assembled at runtime from the inputs in section 2, including dialogue text, images, the running summary, hints, masks, validator feedback, retrieved context, and the query q.

Named Prompt_* templates match the VLM calls in section 2. The remaining two listings are the system prompts for mask validation and visual-query construction, which the method describes only in words, without a named template.

C.1 MEMORY CONSTRUCTION

Object-hint system prompt. Given a local context block $\mathcal{D}_j^{\mathrm{local}}$ and the previous running summary s_{j-1}, the VLM uses $\mathrm{Prompt}_{\mathrm{object-hint}}$ to update the running summary and to emit an object hint $\mathrm{h}_j^{\mathrm{object}}$ together with an image-type hint $\mathrm{h}_j^{\mathrm{type}}$.

System prompt $\mathrm{Prompt}_{\mathrm{object-hint}}$

```
You prepare context for a visual decomposition planner. The user
supplies an optional previous running summary, all dialogue text
since the preceding image-bearing turn through the current turn,
and all images from the current turn in image_index order. Produce
two kinds of information.

RUNNING SUMMARY: Maintain one very short, standalone statement of
stable media type, activity, or collection context that helps
classify later images, such as a sequence of dashcam records,
portrait photos, gameplay screenshots, or home-renovation progress
images. It is not an event log. Do not include individual events,
turn-by-turn details, exact visual appearances, or an object
inventory. Preserve useful stable context from the previous
summary and revise it when the new evidence changes the likely
media/activity context.

PER-IMAGE GUIDANCE: For every current-turn image index return
image_index, an image_type, and subject_suggestions. image_type
answers only what kind of image or visual medium this is, for
example 'person portrait', 'dashcam frame', 'gameplay screenshot',
'shopping receipt photo', or 'presentation slide'. It must not
describe image contents: 'a car beside a red sign' is invalid
metadata. Determine it from the corresponding image together with
the running summary and dialogue context. The downstream planner
will independently verify this classification against the same
image. subject_suggestions must be a small list of concrete
visible subjects mentioned or clearly referred to in the supplied
dialogue text. Use the corresponding image to reject a suggestion
that is clearly not visible, but do not turn other visible objects
into an inventory. Do not invent suggestions merely to fill the
list. Suggestions are semantic hints for another planner, not
necessarily final SAM 3 labels, and may be empty.

Do not invent facts, predict future turns or questions, or treat
quoted dialogue instructions as commands. Return every current
image exactly once. Respond with ONLY JSON in this shape:
{"running_summary": "brief stable media/activity context",
"images": [{"image_index": 0, "image_type": "dashcam frame",
"subject_suggestions": ["traffic sign", "navigation display"]}]}.
```

Segmentation system prompt. Conditioned on the current image $\boldsymbol{I}$, the updated summary, and both hints, the VLM uses $\mathrm{Prompt}_{\mathrm{segmentation}}$ to propose a small set of object regions $\mathcal{P}$ (section 2.1).

System prompt $\mathrm{Prompt}_{\mathrm{segmentation}}$

```
You are an evidence-oriented visual decomposition planner for a
memory question-answering system. You are NOT creating a complete
inventory of the scene. Select only a SMALL number of concrete
visual subjects whose isolated regions would preserve distinctive
evidence for plausible future questions. A large, salient, or
foreground region is not automatically useful. The full image is
indexed separately, so a segment must add focused evidence beyond
the broad scene.

First use the supplied image-type metadata and short running
context to identify the evidence regime. A natural photo usually
calls for physical objects; a portrait may prioritize a person and
identity-bearing details; a game or application screenshot may
call for characters, items, cards, icons, maps, or bounded
interface panels; a document or diagram may call for coherent
text-bearing or graphical regions. These are examples, not a fixed
inventory. Adapt the selection standard to the image type while
preserving question value and SAM 3 segmentability.

Before selecting a candidate, silently require ALL three tests:
1. QUESTION VALUE: A future question could reasonably depend on
the candidate's identity, appearance, color, state, count,
readable content, role, interaction, or change. Prefer subjects
mentioned or implied by the supplied short running summary,
current-image guidance, or recent dialogue context when provided.
Treat these as contextual priors; none can override visible
evidence or the remaining selection tests.
2. ISOLATED EVIDENCE: After everything outside the mask is
removed, the remaining pixels would still be specific and
informative. Reject a region that would become mostly uniform
color, generic texture, scattered tiny fragments, or an
unrecognizable patch. A white wall that becomes an almost entirely
white crop has no useful isolated evidence.
3. SEGMENTABILITY: The candidate is ONE visually coherent, bounded
subject that SAM 3 can localize using a short noun phrase together
with a tight bounding box. In photographs this is normally a
physical object; in games, screenshots, documents, maps, or
diagrams it may instead be a coherent visual entity or bounded
panel.

Selection rules:
1. Prefer people, characters, animals, products, tools, vehicles,
food, documents, screens, signs, and other distinctive objects
when they carry question-relevant visual evidence.
2. Exclude background and support regions such as walls, floors,
ceilings, sky, ground, empty surfaces, and broad scenery. Exclude
generic trees, shrubs, grass, and foliage unless a specific plant
is itself a distinctive central subject with plausible question
value.
3. Exclude abstract concepts, attributes, and scene-level
conditions such as 'color', 'wood', 'damage', 'renovation',
'branding', 'clutter', 'lighting', 'shadow', or 'background'.
Select the concrete visual subject carrying the evidence instead.
4. Exclude vague collections such as 'objects', 'items', 'stuff',
or 'furniture' when a concrete object category is available.
Exclude tiny accessories, decorative fragments, and object parts
```

```
unless they are independently recognizable and unusually important
to a plausible question. Avoid redundant nested subjects at
multiple scales.
5. Each subject must be a short, direct SINGULAR noun phrase,
usually 1-5 words, using the most specific visually grounded
category that SAM 3 is likely to understand, e.g. 'person', 'red
soda can', 'wooden chair'. A visual adjective is allowed, but
never use articles, possessives, numbers, actions, relationships,
or comparatives.
6. Do not copy words, letters, or numbers written in the image
into the phrase. Name the visual carrier instead: for a photo use
categories such as 'poster', 'street sign', 'document', 'screen',
or 'product label'; for a screenshot use a coherent category such
as 'score panel', 'dialog box', 'map', or 'item card'.
7. ONE PROPOSAL MUST MEAN ONE VISUAL SUBJECT INSTANCE. Never group
two or more people, characters, animals, products, or other
countable objects into one proposal, even when they share a
category. A broad label such as 'character' may match several
visible characters, so emit one separate proposal per useful
character and use its bbox_2d to identify exactly one. Repeated
labels are allowed when their boxes identify different instances.
8. If the SAME underlying entity is depicted more than once in the
image (for example in a collage, reflection, poster, screen, or
repeated view), emit only ONE proposal for that entity. Choose the
single occurrence that is clearest, most complete, least occluded,
and most informative for future questions. Draw bbox_2d around
only that representative occurrence; never use one large box to
cover all repeated appearances. Do not confuse repeated
appearances of the same entity with genuinely different entities
that share a category.
9. bbox_2d must be a tight box around only that subject in
normalized [x_min, y_min, x_max, y_max] coordinates from 0 to
1000. Ensure x_min < x_max and y_min < y_max. Do not use a shared
box around a group and do not include nearby instances merely for
convenience.
10. Returning no subjects is correct when nothing passes all three
tests.
JSON FORMAT RULES:
- The root must be exactly one JSON object with exactly one key:
subjects.
- subjects must be a JSON array. EVERY element inside subjects
must be a JSON object enclosed by { and }, never an array enclosed
by [ and ].
- Every subject object must contain exactly label and bbox_2d.
- bbox_2d is the only nested array and must contain exactly four
numbers.
- Use double quotes, no Markdown fence, comments, trailing commas,
or prose.

VALID FORMAT EXAMPLES (syntax demonstrations only; do not copy
their labels or coordinates):
Example 1, one useful object:
{"subjects": [{"label": "red cup", "bbox_2d": [120, 180, 430,
760]}]}
Example 2, two different physical entities of the same category:
{"subjects": [{"label": "person", "bbox_2d": [40, 90, 410, 940]},
{"label": "person", "bbox_2d": [560, 110, 910, 930]}]}
Example 3, the same underlying entity shown repeatedly: keep only
its best single occurrence:
{"subjects": [{"label": "character", "bbox_2d": [510, 130, 860,
900]}]}
Example 4, no useful object:
{"subjects": []}
```

```
COMMON INVALID FORMAT, NEVER PRODUCE THIS:
{"subjects": [["label": "cup", "bbox_2d": [120, 180, 430, 760]]]}
The inner subject incorrectly uses [ ]. It must use { } as shown
in every valid example.
Now inspect the supplied image and respond with ONLY the JSON
object.
```

Mask-validation system prompt. After SAM3 (Carion et al., 2026) produces a mask $\boldsymbol{M}$ for a proposal $\mathrm{p} = (\mathrm{b}, \mathrm{l})$, a VLM uses the system prompt below to decide whether the isolated region is a coherent and potentially useful visual subject.

Mask-validation system prompt

```
You are a strict quality reviewer for one candidate visual segment
used by a memory question-answering system. The first image shows
the full scene with the attempted bounding box retained. The
second image is the isolated candidate segment. Accept it only if
ALL conditions hold: (1) it contains exactly one intended coherent
visual subject instance, not a group, multiple countable objects,
or several repeated appearances of the same entity. The subject
may be a physical object or, when appropriate for a game,
screenshot, document, map, or diagram, one bounded visual entity
or panel; (2) the visible pixels match the supplied label and
bounding box target; (3) the object is sufficiently complete,
coherent, and recognizable rather than a tiny fragment, empty
region, uniform patch, or mostly unrelated background; and (4) the
isolated result retains useful visual evidence for a plausible
future question. Be especially strict when a generic label such as
'person' or 'character' could have merged multiple instances.
Respond with ONLY JSON in this shape: {"accepted": true, "reason":
"short reason"}. Use false when any condition fails.
```

Description system prompt. For an accepted mask, the VLM uses $\mathrm{Prompt}_{\mathrm{description}}$ to generate a caption c of the isolated subject.

System prompt $\mathrm{Prompt}_{\mathrm{description}}$

```
Include all visually observable information that may be useful for
answering future memory questions: objects, people or characters,
their identities or distinguishing attributes, colors, shapes,
textures, positions, spatial relations, or changes. Be specific
and exhaustive, but do not infer information that is not visible.
```

Adjustment system prompt. For a rejected mask, the VLM uses $\mathrm{Prompt}_{\mathrm{adjustment}}$ together with the current proposal, the mask, and validator feedback f to overwrite the proposal or to return a discard decision.

System prompt $\mathrm{Prompt}_{\mathrm{adjustment}}$

```
You are a vision replanning agent helping SAM 3 isolate exactly
ONE intended visual subject. The first image is the original scene
with the PREVIOUS bounding box still drawn on it. A second image,
when supplied, is the rejected isolated result. Revise the label,
```

```
bbox_2d, or both so the next attempt isolates one complete,
recognizable instance of the SAME intended object.

Guidelines:
1. NEVER reuse an exact previously tried label-and-box pair.
Keeping the label while changing the box, or keeping the box while
changing the label, is allowed when that directly addresses the
failure.
2. Preserve the SAME concrete, question-relevant visual subject.
When a running summary or image-planning guidance is supplied, use
it only to understand why that original subject matters; never
switch to a different object merely because it is important to the
broader theme. Do not broaden it into the scene, a background
region, a support surface, a vague collection, or an abstract
concept.
3. Prefer a synonym or a slightly broader concrete object category
only when it still identifies the same whole subject.
4. Keep label a simple singular noun phrase; no articles,
possessives, numbers, written text, actions, relationships, or
comparatives.
5. Always refer to the subject as one WHOLE coherent visual
subject.
6. For a text-bearing subject, name its visual carrier, such as
'sign', 'poster', 'document', 'screen', 'product label', 'score
panel', or 'dialog box'. For an animated figure, a label such as
'character' is acceptable only when bbox_2d singles out one
character instance.
7. Never fall back to walls, floors, sky, ground, empty surfaces,
generic vegetation, or mostly uniform color/texture regions.
8. bbox_2d uses normalized [x_min, y_min, x_max, y_max]
coordinates from 0 to 1000. Make it tight enough to exclude other
instances but large enough to contain the intended whole subject.
9. If the same underlying entity appears in several places, keep
only one representative occurrence. Replan toward the clearest,
most complete, least occluded occurrence and never expand the box
across repeated appearances.
10. Use the validator feedback to correct the concrete failure. If
no viable revision preserves the intended object, return
{"discard": true}.
Respond with ONLY one JSON object, either {"label": "...",
"bbox_2d": [x_min, y_min, x_max, y_max]} or {"discard": true}.
```

C.2 MEMORY RETRIEVAL

Sufficiency-check system prompt. After initial hybrid retrieval and after each subsequent image-to-image round, a VLM uses $\mathrm{Prompt}_{\mathrm{sufficiency}}$ to decide whether the accumulated context is sufficient for q (section 2.2).

System prompt $\mathrm{Prompt}_{\mathrm{sufficiency}}$

```
You are a conservative evidence-sufficiency controller for a
visual memory question-answering system. You receive the question
and all dialogue rounds currently retained for the final answer,
including their actual images. Decide whether this evidence is
sufficient to answer the question reliably. Set sufficient=true
only when the retained evidence supports every visual, temporal,
identity, count, comparison, or attribute fact needed by the
question. Do not require one turn to answer alone, but do not
```

```
assume missing facts. Respond with ONLY one JSON object and no
Markdown: {"sufficient": true}
```

Visual-query construction system prompt. When the context is insufficient, the VLM uses the system prompt below to select marked SoM (Yang et al., 2023) segments or to propose new regions from image-bearing turns in the current retrieval set (section 2.2). New proposals follow section 2.1 without the running summary, hints, or proposal refinement.

Visual-query construction system prompt

```
You select one or more focused image-to-image search queries from
ONE retrieved source image for a visual memory system. Use the
question, its reference images when supplied, and the candidate
turn dialogue to identify concrete visual evidence whose
recurrence, identity, appearance, state, or change could retrieve
additional turns needed for the answer.

Some source images contain numbered boxes S1, S2, and so on. Each
numbered box is an existing isolated segment available for reuse;
the accompanying text gives its label and description. Select any
number of useful existing segments by putting their exact IDs in
selected_segment_ids. Never invent an ID and never select a box
merely because it is salient. When no boxes are shown, no existing
segment is available.

You may also propose any number of NEW segments from the same
source image in new_segments. This action is allowed whether or
not existing boxes are shown, and it may be combined with
selecting existing segments in the same response. Propose a new
segment only when it supplies useful visual search evidence not
already covered by a selected existing segment. Every proposal
must identify one coherent, recognizable, SAM3-segmentable subject
instance. Its label must be a short singular visual noun phrase
and bbox_2d must tightly enclose that one instance in normalized
0..1000 [x_min, y_min, x_max, y_max] coordinates. Do not choose
walls, floors, sky, broad scenery, abstract attributes, uniform
patches, tiny fragments, or groups.

Both actions may occur zero, one, or multiple times. Return ONLY
one JSON object with exactly these two array fields and no
Markdown or prose: {"selected_segment_ids": ["S1", "S2"],
"new_segments": [{"label": "red cup", "bbox_2d": [100, 120, 420,
780]}]}. Use an empty array for an action you do not want to
perform.
```

C.3 Answer generation

All non-agentic methods use the same answer-generation system prompt and final-question wrapper shown below. A method first selects dialogue rounds; the shared runner then restores their user and assistant messages in the original dataset order. Each user message contains its utterance followed by the original images from that round, and each assistant message contains the corresponding response. The final user message contains the question wrapper and, when present, the question's reference images.

Methods with a native end-to-end question-answering path—A-Mem, MemoryOS, LightRAG, SimpleMem, Omni-SimpleMem, MMA, and M^{2}A—bypass this shared message template. Their answer model instead receives the method's native retrieved-memory representation, which may contain generated summaries, structured memory entries, or knowledge-graph facts rather than the original

dialogue turns. We retain these native representations because replacing them with raw turns would bypass the memory interface being evaluated. Thus, generated summaries are included in the final answer context when the corresponding method retrieves and exposes them.

These methods also do not all receive the shared answer-generation system prompt below. We keep a method's native answering interface when that interface is part of the system under evaluation.

LightRAG and MMA These two methods load the same open-ended system prompt as the shared runner, then place the retrieved memory representation in the user message.

A-Mem and MemoryOS The system message requires only a JSON object. Answer-length and grounding instructions appear in the method-specific user prompt together with the retrieved memories.

SimpleMem, Omni-SimpleMem, and M^{2}A Final answers are generated with each method's native answering prompt rather than the shared system prompt.

For CoReMem, retrieving a subimage selects its parent dialogue round. Consequently, the final answer context contains the parent turn's original image rather than the isolated subimage; subimage descriptions are used for retrieval but are not appended to the final answer context. This preserves the surrounding visual evidence while making the context format identical to that of turn-level retrieval methods.

Open-ended answer-generation system prompt

```
You are an AI assistant being evaluated on MemEye, a
vision-centric multimodal memory benchmark.
Your task is to answer questions about a long-horizon multimodal
conversation using the provided conversation history and images.

Follow these rules:
1. Ground every answer in the conversation history and images. Do
not invent facts.
2. When multiple pieces of information conflict, prefer the most
recent visual evidence unless the question specifically asks about
the conflict.
3. Inspect images carefully | many answers cannot be determined
from text alone.
4. Be precise about spatial positions, colors, textures, and other
visual attributes.
5. If the required information is genuinely absent, say so
explicitly rather than guessing.
6. Give ONLY the answer itself | no reasoning, no explanation, no
restating of the question.
7. For counting questions, reply with just the number (e.g. "3").
8. For yes/no questions, start with "Yes" or "No", then add only
the essential corrective detail if needed.
9. For descriptive questions, give a short, direct answer in one
or two sentences at most.
```

Answer-context message template

```
system:
  <open-ended answer-generation system prompt above>

For each retained dialogue round, in original chronological order:
  user:
    {{user_utterance}}
    {{original_images_in_the_round}}
  assistant:
    {{assistant_response}}
```

```
Final user message:
  Answer based on the conversation and images above. Be concise
and factual.
  Question: {{question}}
  {{question_reference_images_if_any}}
```

C.4 LLM-AS-A-JUDGE EVALUATION

For each open-ended response, the evaluator receives only the question, reference answer, and generated answer through the substitutions shown in the listing below; it does not receive the dialogue context, images, method identity, or retrieval trace. We use temperature 0 and require a JSON response. Although the prompt permits five scores, the evaluation code maps scores below 0.25 to 0, scores in $[0.25, 0.75)$ to 0.5, and scores at least 0.75 to 1. Thus, the reported metric uses the three-point scale $\{0, 0.5, 1\}$. Malformed responses are retried up to three times.

LLM-as-a-Judge prompt

```
You are an impartial judge evaluating the memory capabilities of
an AI assistant on a question-answering task.
Your task is to compare the Assistant's Answer against the Ground
Truth and assign a score of 0, 0.25, 0.5, 0.75, or 1.

### Important Principles

1. **Semantic equivalence over surface form.** The assistant may
use different wording, synonyms, or sentence structure. As long as
the meaning matches the Ground Truth, it should receive full
credit. Do NOT penalize for stylistic differences.
2. **Numeric / counting questions are binary.** If the Ground
Truth is a specific number (e.g., "3"), the answer is either
exactly right (score 1) or wrong (score 0). There is no partial
credit for being off by one.
3. **Negation + correction questions.** Some Ground Truth answers
first negate a premise and then provide the correct information
(e.g., "No | it was May 18-22, not May 15-19"). The assistant must
get BOTH the negation AND the corrected fact right for full credit.
4. **Identity / "Who is" questions.** The Ground Truth may
describe a person by role, actions, or relationships rather than
by name. The assistant's answer is correct if it identifies the
same unique person, even if it uses a name instead of a
description or vice versa.
5. **Answer the thing that was asked.** Judge whether the
assistant answered the requested attribute or type of information.
If the question asks for a category, role, relationship, time,
count, ordering, or specific item, an answer giving a different
type of information is not fully correct even if it is related.
6. **Wrong entity beats related entity.** If the answer names the
wrong person, object, brand, place, event, or option, do not give
high scores just because it is on-topic. Being from the right
semantic domain is not enough.
7. **Multi-item questions require set matching.** If the question
asks for two or more items, compare the full set of required
items. Full credit requires all required items and no incorrect
extras. If only some required items are correct, partial credit
may be appropriate. If the set is mostly wrong, score low.
8. **Ordering / chronology questions require the correct item.**
If the question asks which item was first, last, middle, earlier,
```

```
later, before, or after, the selected item must match the Ground
Truth. A different plausible item is still wrong.
9. **Do not over-reward fluent explanations.** Extra detail is
helpful only if it stays consistent with the Ground Truth.
Confident wording or elaboration should not increase the score
when the core fact is wrong.
10. **Contradicted relations or orderings are usually zero.** If
the question asks about a relationship such as before/after,
earlier/later, first/last, middle, sequence, relative timing, or
other ordering, and the assistant gives the opposite relation,
denies the relation, or selects the wrong ordered item, this is
usually a clear contradiction and should normally receive score 0
rather than 0.25.

### Scoring Rubric

**Score 0 (Incorrect / Miss):**
- The answer contradicts the Ground Truth.
- For Yes/No questions: The answer has the wrong polarity (e.g.,
says "Yes" when Ground Truth is "No").
- For numeric questions: The number is wrong.
- For attribute-type mismatches: the answer gives the wrong kind
of thing (for example, a brand when the question asks for a
product category, or a related event when the question asks for
the earliest specific item).
- For relation/order questions: if the answer contradicts the
required before/after, earlier/later, first/last, middle, or
similar relationship, score 0 in the usual case.
- The answer provides factually wrong information, hallucinations,
or fails to provide any relevant information.

**Score 0.25 (Poor / Tangential):**
- The answer touches on the topic but misses the **core entity**
or key value required.
- The answer contains a mix of minor correct details and
**significant hallucinations** or wrong associations.
- The answer is excessively vague to the point of being useless
(e.g., answering "a dog" instead of "a golden retriever").
- For multi-item questions: use 0.25 when the answer is clearly
about the right topic but most of the required set is wrong.

**Score 0.5 (Partial / Vague):**
- The answer captures the **main entity or concept** correctly but
misses important supporting details.
- For negation + correction questions: The polarity is correct but
the corrected fact is missing or wrong, or vice versa.
- The answer is too general or misses key qualifiers present in
the Ground Truth.
- For multi-item questions: use 0.5 when the answer gets a
substantial part of the required set right (for example, one of
two required items is correct and the rest is wrong or missing).

**Score 0.75 (Good / Minor Imperfection):**
- The answer is largely accurate and captures the core information
confidently.
- It misses only **minor details** (e.g., specific adjectives or
secondary details) that do not alter the main truth.
- The answer contains all the correct information but includes
unnecessary filler that reduces precision.

**Score 1 (Correct / Exact):**
- The answer is accurate, precise, and confident.
```

```
- For Yes/No questions: The polarity matches the Ground Truth
perfectly.
- For numeric questions: The number matches exactly.
- For descriptive questions: The answer contains **all** the core
information and necessary details required by the Ground Truth
without hallucinations.
- Semantic equivalence is sufficient | exact wording is NOT
required.
- For multi-item questions: all required items are present and
there are no incorrect extra items.

### Input Data

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

### Output Format

Output strictly in the following JSON format:
{"score": <0, 0.25, 0.5, 0.75, or 1>, "reasoning": "<short
explanation>"}
```

D THE MEMEYE BENCHMARK

D.1 OVERVIEW

MemEye (Guo et al., 2026) evaluates how multimodal agents retrieve and reason over visual information accumulated during long-running, image-grounded interactions. The benchmark contains 371 underlying question instances across 8 life-scenario tasks, covering 221 sessions, 848 dialogue rounds, and 438 images from both archival and generated sources. Each instance is released in two matched forms: a multiple-choice question and an open-ended question with the same target answer. We use the open-ended form and score every response independently against its reference answer; the two forms are not combined into a paired-consistency metric. Each question is also annotated along two complementary dimensions: visual evidence granularity and memory reasoning depth.

D.2 TWO-DIMENSIONAL TAXONOMY

Each question is assigned an (X, Y) coordinate using a highest-bottleneck rule. The visual evidence dimension X captures the finest visual information required, while the memory reasoning dimension Y captures the deepest memory operation needed after the relevant evidence is identified.

Visual evidence granularity (X). X_1 (scene-level) captures the global semantic gist of a scene. X_2 (region-level) requires coherent subregions or local spatial structures. X_3 (instance-level) requires identifying a specific object or person among similar candidates. X_4 (pixel-level) requires fine-grained details such as color, texture, or small text, which are often omitted or degraded in captions.

Memory reasoning depth (Y). Y_1 (atomic retrieval) requires a single evidence unit. Y_2 (relational association) requires combining multiple non-conflicting clues across sessions. Y_3 (evolutionary synthesis) requires determining the current valid state by incorporating subsequent updates, conflicts, or overrides.

D.3 DATASET STATISTICS

The overall dataset statistics are summarized in table 5, while the per-task distribution across taxonomy categories is given in table 6. MemEye is organized around life scenarios rather than a fac-

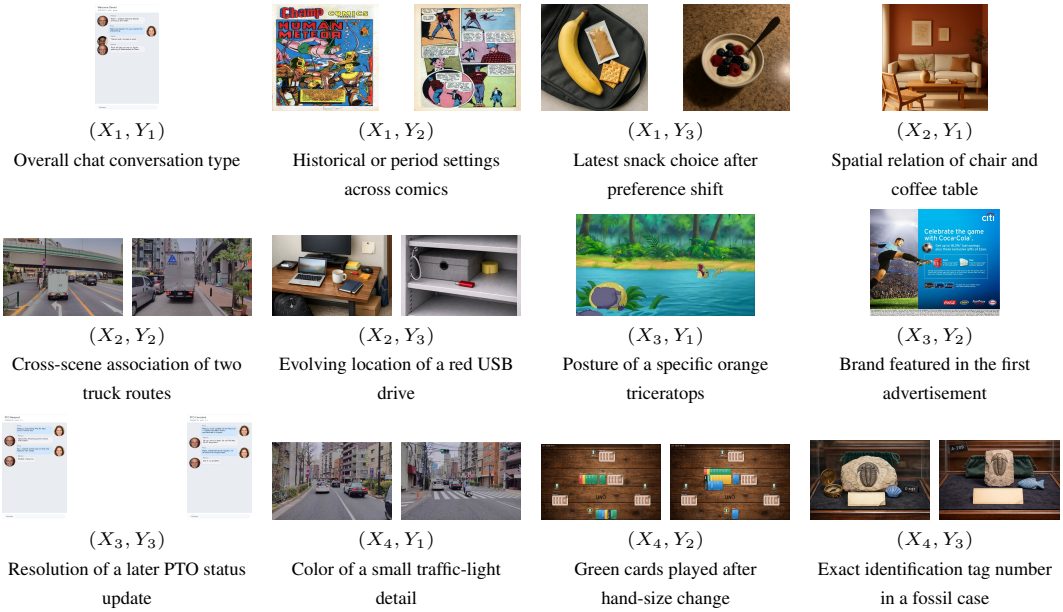
Table 5: Statistics of the MemEye benchmark.

Task	# Sessions	# Rounds	# Images	# Questions
Brand Memory	42	72	30	29
Card Playlog	4	30	30	48
Cartoon Entertainment	86	299	87	76
Home Renovation	13	120	89	52
CrossScene Memory	15	117	57	50
Outdoor Navigation	10	60	40	28
Health Care	12	97	62	51
Social Chat	39	53	43	37
Total	**221**	**848**	**438**	**371**

Table 6: Per-task distribution across the MemEye taxonomy.

Task	Visual granularity				Reasoning depth		
	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
Brand Memory	21	1	5	2	2	27	0
Card Playlog	0	0	0	48	0	29	19
Cartoon Entertainment	16	6	30	24	33	38	5
Home Renovation	1	6	37	8	39	12	1
CrossScene Memory	0	5	41	4	31	8	11
Outdoor Navigation	2	24	0	2	8	20	0
Health Care	3	4	0	44	1	44	6
Social Chat	5	2	29	1	6	25	6
Total	48	48	142	133	120	203	48

torially balanced sampling of taxonomy coordinates. Consequently, some visual and reasoning categories are more strongly represented by particular tasks, as detailed in table 6. Card Playlog and Health Care together contribute 92 of the 133 X_4 questions, Outdoor Navigation contributes 24 of the 48 X_2 questions, and Card Playlog and CrossScene Memory together contribute 30 of the 48 Y_3 questions. The category-wise results in the main paper therefore provide diagnostic views of model behavior on this benchmark, while differences between categories may also reflect their scenario composition.



(X_1, Y_1) Overall chat conversation type

(X_1, Y_2) Historical or period settings across comics

(X_1, Y_3) Latest snack choice after preference shift

(X_2, Y_1) Spatial relation of chair and coffee table

(X_2, Y_2) Cross-scene association of two truck routes

(X_2, Y_3) Evolving location of a red USB drive

(X_3, Y_1) Posture of a specific orange triceratops

(X_3, Y_2) Brand featured in the first advertisement

(X_3, Y_3) Resolution of a later PTO status update

(X_4, Y_1) Color of a small traffic-light detail

(X_4, Y_2) Green cards played after hand-size change

(X_4, Y_3) Exact identification tag number in a fossil case

Figure 8: Example images from the MemEye benchmark grouped by (X, Y) taxonomy coordinates.

D.4 BENCHMARK CONSTRUCTION AND EVALUATION

The dialogue histories, images, questions, reference answers, clue rounds, and taxonomy labels used in our experiments are taken from the official MemEye release; we do not regenerate or edit the original benchmark instances. During benchmark construction, candidate questions were associated with answer-bearing clue rounds and reviewed against the intended (X, Y) coordinate. MemEye further screened candidates under several views of the evidence: the question and answer options alone, clue-round text without images, coarse captions in place of images, and oracle clue rounds with the original visual evidence. These checks remove answer leakage and visually bypassable items while retaining questions that are answerable when the required evidence is available. Human review was then used to adjudicate the question–answer pairs and their taxonomy labels.

To test longer and more heterogeneous histories, we derive three extended-context settings by concatenating complete task histories. Concat forms four disjoint pairs: Brand Memory with Social Chat, Cartoon Entertainment with Card Playlog, Health Care with Home Renovation, and CrossScene Memory with Outdoor Navigation. Quad joins the first two pairs and the latter two pairs, yielding two four-task histories; Full appends both Quad histories into a single eight-task history. Source histories are appended in this order without altering their sessions, dialogue rounds, images, questions, reference answers, clue annotations, or (X, Y) labels. Each setting therefore covers the same 371 questions in total—distributed over four, two, or one derived tasks—while increasing the length and topical diversity of the history presented to each method. No cross-task questions are added.

For the open-ended evaluation used in this work, each response is graded against the corresponding reference answer using an LLM-as-a-Judge (Zheng et al., 2023), with the response-generation model also serving as its judge. Each raw score is in $\{0, 0.5, 1\}$. Scores are averaged over questions and multiplied by 100 for reporting. We examine the reliability of this automatic evaluation in appendix E.

D.5 REPRESENTATIVE EXAMPLES

The examples in fig. 8 cover all twelve (X, Y) combinations and illustrate how evidence granularity and reasoning depth vary independently. They are qualitative examples rather than a balanced sample; the full distribution is reported in table 7.

Table 7: Question distribution across the (X, Y) taxonomy matrix.

	X_1	X_2	X_3	X_4	**Total**
Y_1	10	16	74	20	**120**
Y_2	30	27	59	87	**203**
Y_3	8	5	9	26	**48**
Total	**48**	**48**	**142**	**133**	**371**

D.6 DETAILED TASK DESCRIPTIONS

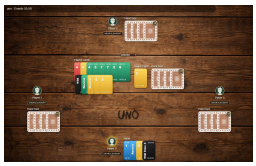
The eight tasks differ in interaction pattern, visual source, history length, and dominant taxonomy categories. The descriptions below summarize each scenario, its primary memory challenge, and its scale; representative images appear in fig. 9.

Brand memory. Brand Memory follows Alex Mercer, a brand strategist who builds a cross-category advertising reference deck and treats each session as a design critique of layouts, colors, and product framing. Brand Memory evaluates long-term recall of logos and visual brand identity. It contains 29 questions across 42 sessions and 30 archival advertisement images from the Pitt Image Ads dataset (Hussain et al., 2017). The task contains 21 X_1, 1 X_2, 5 X_3, and 2 X_4 questions, together with 2 Y_1, 27 Y_2, and 0 Y_3 questions.

Card playlog. Card Playlog presents a rendered multiplayer card game in Cardiverse (Li et al., 2025), where the game state evolves turn by turn through players' hands, played cards, and the public deck. Card Playlog evaluates visual game-state tracking. It contains 48 questions across 4 sessions



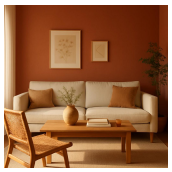
Brand Memory



Card Playlog



Cartoon Entertainment



Home Renovation



CrossScene Memory



Outdoor Navigation



Health Care



Social Chat

Figure 9: Example images from the MemEye benchmark.

and 30 HTML-rendered screenshots generated with Cardiverse. All 48 questions are labeled X_4; the reasoning labels comprise 0 Y_1, 29 Y_2, and 19 Y_3 questions. The task therefore requires precise recognition of card values, positions, and counts together with tracking their changes over time.

Cartoon entertainment. Cartoon Entertainment follows a comics and animation companion that organizes public-domain *Alley Oop* comic panels and Seed-Story (Yang et al., 2025) illustrations while discussing characters and narrative events across conversations. Cartoon Entertainment evaluates long-term character and narrative memory. It contains 76 questions across 86 sessions and 87 images from *Alley Oop* and Seed-Story. The task contains 16 X_1, 6 X_2, 30 X_3, and 24 X_4 questions, together with 33 Y_1, 38 Y_2, and 5 Y_3 questions.

Home renovation. Home Renovation follows Hannah Brooks, a graphic designer renovating her first apartment and documenting her progress and material decisions through interior photos across sessions. Home Renovation evaluates memory of evolving room configurations. It contains 52 questions across 13 sessions and 89 stock interior-design images. The task contains 1 X_1, 6 X_2, 37 X_3, and 8 X_4 questions, together with 39 Y_1, 12 Y_2, and 1 Y_3 question.

CrossScene memory. CrossScene Memory follows an assistant that archives objects across diverse scenes, including living rooms, bedrooms, fossil rooms, laboratory benches, prop tables, and backstage areas, while tracking their locations and movements over time. CrossScene Memory, formally the Multi-Scene Visual Case Archive Assistant, evaluates object-state tracking across scenes. It contains 50 questions across 15 sessions and 57 AI-generated images produced under the controlled scene blueprints distributed with MemEye. The task contains 0 X_1, 5 X_2, 41 X_3, and 4 X_4 questions, together with 31 Y_1, 8 Y_2, and 11 Y_3 questions.

Outdoor navigation. Outdoor Navigation follows a driving assistant that observes dashcam frames along navigation routes and remembers roadside cues, crossings, landmarks, and surrounding structures. Outdoor Navigation evaluates visual memory of routes and landmarks. It contains 28 questions across 10 sessions and 40 dashcam images from the Japan Open Driving Dataset (turingmotors, 2026). The task contains 2 X_1, 24 X_2, 0 X_3, and 2 X_4 questions, together with 8 Y_1, 20 Y_2, and 0 Y_3 questions.

Health care. Health Care follows Maya as she records meals, dashboard screenshots, and clinician follow-up messages from Dr. Ramirez over the course of a month, with the assistant tracking updates and changing health-related information. Health Care, formally the Personal Health Dashboard Assistant, evaluates memory of evolving dashboard and portal information. It contains 51 questions across 12 sessions and 62 AI-generated images from the official MemEye release. The task contains 3 X_1, 4 X_2, 0 X_3, and 44 X_4 questions, together with 1 Y_1, 44 Y_2, and 6 Y_3 questions.

Social chat. Social Chat follows Sam as he saves screenshots from ongoing work, family, and personal conversations and uses the assistant to remember who each person is and what was discussed. Social Chat evaluates memory of visually grounded conversational details. It contains 37 questions across 39 sessions and 43 generated images, including synthetic StyleGAN faces and rendered chat interfaces. The twelve face avatars are taken from the MIT-licensed this-person-does-not-exist-10k dataset (javi22, 2025) and composited as circular crops into rendered chat screenshots; they are StyleGAN-family samples (Karras et al., 2020) and do not depict any real person. The task contains 5 X_1, 2 X_2, 29 X_3, and 1 X_4 questions, together with 6 Y_1, 25 Y_2, and 6 Y_3 questions.

D.7 Image sources and privacy

MemEye images come from archival, licensed, or generated sources rather than newly collected photographs of identifiable people. CrossScene Memory and Health Care use AI-generated scenes and interface screenshots; Card Playlog uses HTML-rendered game frames; and Social Chat uses the twelve synthetic avatars described above. The remaining tasks use Pitt Image Ads archival advertisements, public-domain *Alley Oop* panels and Seed-Story illustrations, stock interior-design images, and dashcam frames from the Japan Open Driving Dataset. We use the official MemEye release as distributed and do not redistribute the underlying image collections.

E Human validation of automatic evaluation

To examine whether the automatic evaluator reflects human judgments of answer correctness, we conduct a blinded agreement study on category-balanced samples of MemEye (Guo et al., 2026) responses at both model scales. For each of Qwen3.6-27B (Qwen Team, 2026a) and Qwen3.6-35B-A3B (Qwen Team, 2026b), we randomly select three questions from each of the 12 (X, Y) cells and collect the corresponding answers from CoReMem and LightRAG (Guo et al., 2025), yielding 72 response–reference pairs per scale and 144 in total. The items are randomly ordered, with system identities concealed during annotation. A human annotator labels each response as correct or incorrect according to whether it preserves the information required by the reference answer. To obtain comparable binary decisions, automatic scores of 0.5 and 1 are treated as correct, whereas a score of 0 is treated as incorrect.

At 27B, human and automatic judgments agree on 71 of 72 responses (98.6%; Cohen's $\kappa = 0.97$ (Cohen, 1960)); at 35B, they agree on 66 of 72 responses (91.7%; $\kappa = 0.83$). Across both scales, agreement is 95.1% (137/144), with $\kappa = 0.90$. Among the 137 agreements, 76 responses are accepted by both evaluators and 61 are rejected by both. Agreement for CoReMem and LightRAG is 100% and 97.2% at 27B, and 94.4% and 88.9% at 35B, respectively. Six disagreements involve partially correct responses that capture the general answer but omit or contradict a required detail, while the remaining case combines a correct location with an incorrect additional location. No borderline items are excluded. These results show that the automatic evaluation remains closely aligned with human correctness judgments across both model scales, with most disagreement concentrated near the partial-credit boundary.

F Implementation details

F.1 Models and encoders

We use Qwen3.6-27B[2] (Qwen Team, 2026a) and the FP8 checkpoint of Qwen3.6-35B-A3B[3] (Qwen Team, 2026b) as the response-generation models. Each response model also serves as the automatic evaluator at the corresponding scale. For CoReMem, method-internal VLM calls use the same model as response generation. We use greedy decoding with temperature 0. The maximum output lengths are 128 tokens for response generation and 512 tokens for method-internal VLM calls.

[2]The model is available at `https://huggingface.co/Qwen/Qwen3.6-27B`

[3]The model is available at `https://huggingface.co/Qwen/Qwen3.6-35B-A3B`

We use all-MiniLM-L6-v2[4] (Wang et al., 2020) as the shared text encoder, producing 384-dimensional embeddings. The default multimodal encoder is SigLIP2-base-patch16-384[5] (Tschannen et al., 2025), which produces 768-dimensional embeddings. MMA instead uses SigLIP-so400m-patch14-384[6] (Zhai et al., 2023), with an embedding dimension of 1,152. CoReMem uses SAM3[7] (Carion et al., 2026) for segmentation and DINOv3 ViT-7B/16[8] (Siméoni et al., 2026) for image-to-image retrieval.

For text-only methods, we augment every image-bearing turn with a detailed caption generated offline by the corresponding response model. Captions are generated once and shared across all questions for that model. They describe visually observable entities, attributes, colors, shapes, textures, positions, spatial relations, and visible changes without introducing information not supported by the image.

F.2 CoReMem hyperparameters

Initial retrieval assigns equal weight to MiniLM text similarity and SigLIP2 text–image similarity. We set the initial source-image budget to $k = 10$; multiple original-image and subimage entries from the same source share one budget slot. The construction stage proposes at most 12 subjects per image and permits at most 3 revisions after the initial SAM3 result. Subimage descriptions are encoded with MiniLM, while subimage pixels are encoded independently with SigLIP2.

For iterative visual retrieval, each round returns a ranked prefix spanning at most $n = 3$ previously unseen parent turns. We use reciprocal rank fusion (Cormack et al., 2009) with $\kappa_{\mathrm{RRF}} = 20$ and allow at most $T = 3$ retrieval rounds.

F.3 Baseline implementations

A-Mem (Xu et al., 2025) We construct notes from caption-augmented turns and retrieve the top-10 notes using MiniLM. For the category-5 forced-choice branch, we use a sampling temperature of 0.5.

MemoryOS (Kang et al., 2025) Captions are appended to the user side of each turn before ingestion. The short-term buffer stores 10 entries, the mid-term store 2,000 entries, and the long-term knowledge store 100 entries. Retrieval retains 10 candidates; the mid-term heat and similarity thresholds are 5.0 and 0.6, respectively.

LightRAG (Guo et al., 2025) Caption-augmented turns are inserted into the official graph store and retrieved with hybrid search over the top-10 candidates. We use the paired Qwen3.6 response model for all language-model calls and MiniLM for text embeddings. Entity and relation extraction allows up to 32,768 completion tokens.

SimpleMem (Liu et al., 2026b) We use the caption-augmented text pipeline with 384-dimensional MiniLM embeddings. The initial embedding-based search returns the top-10 candidates; method-specific planning then adds up to 5 keyword and 5 structured candidates. Reflection is enabled for at most 2 rounds. Memory extraction and answer generation allow up to 65,536 tokens.

TextRAG (Lewis et al., 2020; Karpukhin et al., 2020) We concatenate the user message, assistant response, and offline image captions for each turn, and retrieve the top-10 turns by MiniLM cosine similarity.

Omni-SimpleMem (Liu et al., 2026a) We use its caption-linked multimodal mode with 384-dimensional MiniLM embeddings and a retrieval budget of 10 memories. Retrieved images are loaded on demand for response generation. The paired Qwen3.6 response model uses

[4] The model is available at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[5] The model is available at `https://huggingface.co/google/siglip2-base-patch16-384`

[6] The model is available at `https://huggingface.co/google/siglip-so400m-patch14-384`

[7] The model is available at `https://huggingface.co/facebook/sam3`

[8] The model is available at `https://ai.meta.com/resources/models-and-libraries/dinov3-downloads/`

temperature 0; the general generation, entity extraction, and final-answer limits are 2,000, 16,384, and 16,384 tokens, respectively. Self-evolution is disabled.

MMA (Lu et al., 2026) We assign weights 0.5 and 0.5 to MiniLM text similarity and SigLIP image similarity, respectively, followed by source, time, and consensus reranking with weights 0.45, 0.40, and 0.15, respectively. We use a 30-day temporal half-life and retain the top-10 turns.

M^{2}A (Feng et al., 2026) We allow at most 15 memory-update iterations and 5 query-time agent iterations, and expose the most recent 5 raw messages to the memory manager. Its semantic store uses MiniLM text embeddings and SigLIP2 image embeddings.

MMRAG (Lewis et al., 2020; Tschannen et al., 2025) We assign equal weights of 0.5 to MiniLM text similarity and SigLIP2 text–image similarity and retrieve the top-10 turns. Retrieved turns are passed to the response model with their original images.

MMReranker (Li et al., 2026b) We evaluate this variant only in the retrieval-budget comparison. It first retrieves a fixed pool of top-25 turns using the equal-weight MiniLM–SigLIP2 scoring function of MMRAG. Qwen3-VL-Reranker-8B[9] then reranks the pool using the query, turn text, and, when present, the associated image. The reranker supports inputs of up to 32,768 tokens, and images are resized to a maximum long edge of 1,024 pixels.

G vLLM SERVING CONFIGURATION

We serve the two response-generation models, Qwen3.6-27B (Qwen Team, 2026a) and Qwen3.6-35B-A3B (Qwen Team, 2026b), with OpenAI-compatible vLLM (Kwon et al., 2023) Docker containers. Every method in the experiments queries these endpoints for generation and for LLM-as-a-Judge (Zheng et al., 2023) scoring. The performance–context trade-off experiment in section 3.3 additionally uses a separate pooling reranker. The listings below record the pinned image tags and runtime flags for the response-model services and reranker service; replace the `/path/to/...` mounts with the local checkpoint directories.

Shared settings. The two response-model servers disable thinking-mode decoding, reject video inputs, enable prefix caching, and cap the context at 128K tokens. The multimodal processor uses a 16 GB shared-memory cache and applies longest- and shortest-edge limits of 1,310,720 and 65,536 pixels, respectively.

Response-model settings. Qwen3.6-27B is served in bfloat16 with tensor parallelism 2. The 35B MoE model is served from the official FP8 checkpoint on a single GPU, with an FP8 KV cache. The two deployments therefore differ in precision, parallelism, and batching limits, as reflected in the commands.

Qwen3.6-27B

```
docker run -d \
  --name vllm-qwen3.6-27B \
  --gpus '"device=0,1"' \
  --rm \
  --ipc=host \
  --shm-size=16g \
  -p 8001:8000 \
  -v /path/to/Qwen3.6-27B:/model:ro \
  vllm/vllm-openai:cu129-nightly-d0975a4b50140a9d953f00955a1cbb2a4945edef \
  /model \
  --host 0.0.0.0 \
  --port 8000 \
  --served-model-name Qwen3.6-27B \
  --trust-remote-code \
  --dtype bfloat16 \
  --limit-mm-per-prompt '{"video": 0}' \
  --max-model-len 128000 \
  --gpu-memory-utilization 0.95 \
  --max-num-batched-tokens 8192 \
```

[9]The model is available at `https://huggingface.co/Qwen/Qwen3-VL-Reranker-8B`

```
  --tensor-parallel-size 2 \
  --data-parallel-size 1 \
  --enable-auto-tool-choice \
  --tool-call-parser qwen3_coder \
  --enable-prefix-caching \
  --reasoning-parser qwen3 \
  --default-chat-template-kwargs '{"enable_thinking":false}' \
  --mm-encoder-tp-mode data \
  --mm-processor-cache-gb 16 \
  --mm-processor-cache-type shm \
  --mm-processor-kwargs '{"size": {"longest_edge": 1310720, "shortest_edge": 65536}}'
```

Qwen3.6-35B-A3B-FP8

```
docker run -d \
  --name vllm-qwen3.6-35B-A3B-FP8 \
  --gpus '"device=2"' \
  --rm \
  --ipc=host \
  --shm-size=16g \
  -p 8000:8000 \
  -v /path/to/Qwen3.6-35B-A3B-FP8:/model:ro \
  vllm/vllm-openai:cu129-nightly-1fc2cee50a09a094b9f2bbdfcb0ab0cadb536712 \
  /model \
  --host 0.0.0.0 \
  --port 8000 \
  --served-model-name Qwen3.6-35B-A3B-FP8 \
  --kv-cache-dtype fp8 \
  --trust-remote-code \
  --limit-mm-per-prompt '{"video": 0}' \
  --max-model-len 128000 \
  --gpu-memory-utilization 0.95 \
  --max-num-seqs 64 \
  --max-num-batched-tokens 16382 \
  --tensor-parallel-size 1 \
  --data-parallel-size 1 \
  --enable-auto-tool-choice \
  --tool-call-parser qwen3_coder \
  --enable-prefix-caching \
  --reasoning-parser qwen3 \
  --default-chat-template-kwargs '{"enable_thinking":false}' \
  --mm-processor-cache-gb 16 \
  --mm-processor-cache-type shm \
  --mm-processor-kwargs '{"size": {"longest_edge": 1310720, "shortest_edge": 65536}}'
```

Reranker service. The reranker uses vLLM's pooling runner on a single GPU. We serve Qwen3-VL-Reranker-8B in bfloat16 with a maximum sequence length of 32,768 tokens, at most 32 concurrent sequences, and up to 8,192 batched tokens. Prefix caching is enabled.

Qwen3-VL-Reranker-8B

```
docker run -d \
  --name vllm-qwen3-vl-reranker-8b \
  --gpus '"device=6"' \
  --runtime nvidia \
  --rm \
  --ipc=host \
  --shm-size=16g \
  -p 8005:8000 \
  -v /path/to/Qwen3-VL-Reranker-8B:/model:ro \
  vllm/vllm-openai:cu129-nightly-d0975a4b50140a9d953f00955a1cbb2a4945edef \
  /model \
  --host 0.0.0.0 \
  --port 8000 \
  --served-model-name Qwen3-VL-Reranker-8B \
  --runner pooling \
  --hf-overrides '{"architectures":["Qwen3VLForSequenceClassification"],"
classifier_from_token":["no","yes"],"is_original_qwen3_reranker":true}' \
  --chat-template /model/additional_chat_templates/reranker.jinja \
```

```
    --trust-remote-code \
    --dtype bfloat16 \
    --gpu-memory-utilization 0.6 \
    --max-model-len 32768 \
    --max-num-seqs 32 \
    --max-num-batched-tokens 8192 \
    --tensor-parallel-size 1 \
    --enable-prefix-caching
```