MEMEYE: A VISUAL-CENTRIC EVALUATION FRAMEWORK FOR MULTIMODAL AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term agent memory is becoming increasingly multimodal, but existing evaluations rarely test whether agents can preserve visual details across long interactions. In many prior benchmarks, questions involving images can still be answered from captions, dialogue, or other textual cues, without relying on the original visual information. Existing evaluations also provide limited coverage of cases where visual states change over time and later evidence updates or replaces earlier observations. We introduce **MemEye**, a benchmark built around two challenges in multimodal memory: preserving visual information at the right level of detail and using that information correctly as memory evolves over time. We capture these challenges with two axes, ranging from scene-level to pixel-level visual evidence, and from single-evidence retrieval to reasoning over changing states. Under this framework, we construct a new benchmark of 1,200 questions across eight life-scenario tasks, with 100 questions per taxonomy cell and ablation-driven validation gates for assessing answerability, shortcut resistance, visual necessity, and reasoning structure. We evaluate 13 memory methods across four vision-language model (VLM) backbones and find that current systems *still struggle with both fine-grained visual retention and reasoning over evolving visual states*. Further analysis shows that failures often come from lost visual details, stale or incomplete retrieval, and difficulty tracking the current visual state.

1 INTRODUCTION

Long-term memory is increasingly important for agents that operate across extended interactions (Zhang et al., 2025; Chhikara et al., 2025). VLMs (Bai et al., 2025; Xie et al., 2024) now allow agents to interact through both text and images, making visual information an important part of what must be remembered. In multimodal conversations, agents need to retain both dialogue history and visual information across sessions. Human visual memory can preserve detailed object information and track how scenes change over time (Brady et al., 2008), but current benchmarks do not clearly test whether VLM-based agents can do the same over long interactions.

Most existing benchmarks (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2026a; Das et al., 2017; Kottur et al., 2019; Liu et al., 2024b; Xue et al., 2025) either study image understanding over short contexts or evaluate long-term memory in settings dominated by text. Recent benchmarks such as Mem-Gallery (Bei et al., 2026) place images across multiple sessions, but many of their visually grounded questions can still be answered from captions, nearby dialogue, or the answer choices alone. As shown in Figure 2, LoCoMo (Maharana et al., 2024), MMRC (Xue et al., 2025), and Mem-Gallery (Bei et al., 2026) show relatively small gains when moving from captions to the original images, indicating that much of the answer-relevant information is still recoverable from text. These benchmarks also rarely make visual state changes themselves central to the task. Updates are often stated in dialogue rather than conveyed through changing images, which makes it hard to evaluate whether an agent can track how a visual state evolves over time.

Two problems remain difficult to evaluate separately. One is whether a memory system can preserve visual information at the right granularity. Captions may capture the overall scene while dropping local layouts, object identities, textures, or other fine-grained cues. The other is whether the system can use information accumulated over time, especially when evidence from different sessions must be linked or when newer observations replace earlier states. Current benchmarks often mix these two

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

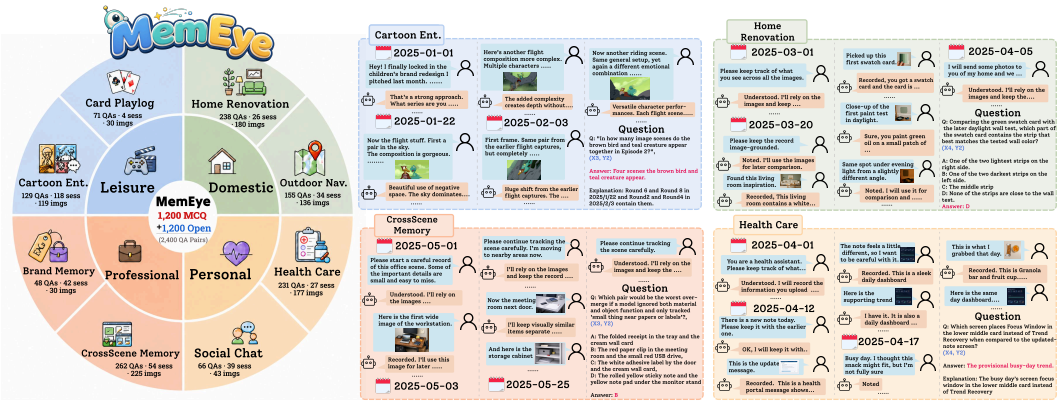


Figure 1: The left sub-figure shows the MemEye dataset overview, with inner rings grouping tasks and outer rings showing statistics. The right sub-figure presents example cases.

sources of difficulty, making it hard to tell whether a failure comes from losing the visual evidence itself or from using the stored evidence incorrectly.

To bridge this gap, we introduce MemEye, a framework that evaluates multimodal agent memory along two orthogonal dimensions. The first dimension, ***visual evidence granularity***, ranges from scene-level evidence to pixel-level evidence and measures whether memory systems can preserve the visual details needed to answer a question. The second dimension, ***memory reasoning depth***, ranges from atomic retrieval to evolutionary synthesis and measures whether memory systems can reason over retrieved evidence to answer a question. Based on this framework, we construct a benchmark with 1,200 mirrored multiple-choice and open-ended questions across eight life-scenario tasks, balanced across the 12 taxonomy cells with 100 questions per cell, as shown in Figure 1. Each question is annotated with its visual-evidence granularity and memory-reasoning depth, and filtered to reduce textual shortcut cases. Figure 2 shows that MemEye gains more from access to the original images than prior long-term memory benchmarks, suggesting that its questions rely more heavily on visual information that is not preserved by captions alone.

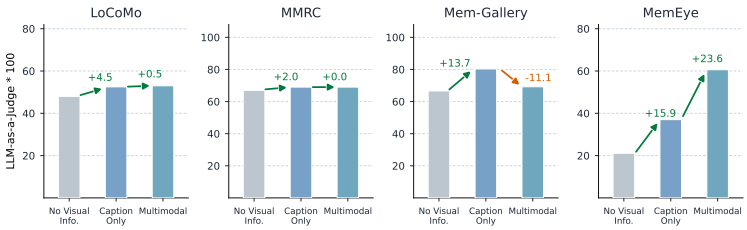


Figure 2: Measuring visual information necessity across long-term memory benchmarks. We compare three settings with the same textual question: *No Visual Info.* removes images, *Caption Only* replaces images with captions, and *Multimodal* provides the original images. MemEye exhibits stronger visual irreplaceability than LoCoMo (Maharana et al., 2024), MMRC (Xue et al., 2025), and Mem-Gallery (Bei et al., 2026). Details are in Appendix D.3.

Using this benchmark, we evaluate 13 memory methods across four vision-language model backbones. Current systems remain far from reliable long-term visual memory. We identify a trade-off: text-based memory can help organize state transitions and updates, but often loses fine-grained visual details during abstraction. In contrast, native image memory preserves visual evidence more directly, but struggles to identify which visual state remains valid over time. Furthermore, cross-topic scaling shows that specialized memory mechanisms become more important as history length and thematic diversity grow. Together, these findings suggest that effective multimodal memory must preserve visual details, track temporal validity, and select the right evidence over long histories. In summary, our main contributions are:

- **A Multi-dimensional Evaluation Framework.** We propose MemEye, a novel framework that categorizes multimodal memory challenges along two orthogonal axes: visual evidence granularity, ranging from scene-level to pixel-level, and memory reasoning depth, ranging from atomic retrieval to evolutionary synthesis.

- **A Vision-Centric Long-Term Memory Benchmark.** We introduce a rigorous benchmark with 1,200 questions across real-world scenarios, balanced across 12 taxonomy cells, to test whether multimodal agents can preserve and reason over irreplaceable visual evidence.
- **Comprehensive Evaluation and Empirical Insights.** We comprehensively evaluate existing methods and reveal a key trade-off: text-based memory helps manage state changes but loses fine-grained visual details, while image-based memory preserves visual evidence but struggles with temporal validity.

2 Related Work

Agent Memory Systems. Long-term memory management is a central design problem for deployed agents (Zhang et al., 2025; Hu et al., 2026b). Prior work has explored memory mechanisms for computer-use and interactive agents (Shi et al., 2025), including textual memory systems with explicit memory writing, updating, and maintenance procedures (Xu et al., 2026; Kang et al., 2025; Chhikara et al., 2025). These methods improve an agent's ability to store and reuse past information, but they primarily operate over textual memories or text abstractions of prior experience.

Recent multimodal memory methods extend this line of work by retaining or retrieving visual experience, including MIRIX (Wang & Chen, 2025), MMA (Lu et al., 2026), and M2A (Feng et al., 2026). These systems make different architectural trade-offs among coverage, retrieval selectivity, abstraction, and revision. However, existing evaluations often report end-task performance without isolating which memory operation fails. A memory system may discard perceptual details during captioning or memory writing, retrieve a semantically relevant but temporally invalid clue, or fail to synthesize the valid state even when relevant evidence is available. MemEye is designed to evaluate these mechanisms rather than only compare aggregate accuracy, exposing when a memory architecture loses visual evidence, selects stale evidence, or fails to recover the valid memory state.

Table 1: Comparison with multimodal conversational benchmarks.

Benchmark	Caption-Proof	Long Memory	Fine Visual	Cross-Context	State Revision	Visual-State
VisDial (Das et al., 2017)	✗	✗	✓	✓	✗	✗
CLEVR-dialog (Kottur et al., 2019)	✗	✗	✓	✓	✗	✗
LoCoMo (Maharana et al., 2024)	✗	✓	✗	✓	●	✗
MMDU (Liu et al., 2024b)	✗	✗	✓	●	✗	✗
ConvBench (Liu et al., 2024a)	✗	✗	✓	✓	✗	✗
MMRC (Xue et al., 2025)	✗	●	✗	✓	✓	✗
MultiVerse (Lee et al., 2025)	✗	✗	✓	✓	✗	✗
Mem-Gallery (Bei et al., 2026)	✗	✓	●	✓	●	✗
MemEye	✓	✓	✓	✓	✓	✓

* ✓ = substantial ($\geq$20%), ● = limited (5–20%), ✗ = negligible ($<$5%) coverage under MemEye labeling. **Caption-Proof**: verifies native images cannot be replaced by captions. **Long Memory**: requires extended conversational memory. **Fine Visual**: X_3–X_4 visual bottlenecks (instance binding, pixel attributes, OCR). **Cross-Context**: Y_2 associations across turns, sessions, or modalities. **State Revision**: Y_3 reasoning under updates, conflicts, or overrides. **Visual-State**: joint high-visual, evolving-memory region.

Memory Benchmarks for Long-Horizon Multimodal Agents. Long-horizon agent benchmarks increasingly evaluate whether systems can retain information across extended interactions. Text-centric benchmarks such as LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2025), TwinVoice (Du et al., 2025), and MemoryAgentBench (Hu et al., 2026a) primarily measure whether linguistic facts can be recovered, summarized, or used after many turns. Multimodal benchmarks such as MMDU (Liu et al., 2024b), ATM-Bench (Mei et al., 2026), and MMRC (Xue et al., 2025) introduce image information within dialogue, while Mem-Gallery (Bei et al., 2026) further extends this direction to a multi-session multimodal memory setting where images appear throughout the conversation.

The missing object of study is not only another task domain, but the coupled failure mode between visual evidence compression and state-evolving memory use. As shown in Table 1, prior benchmarks rarely ask whether the decisive image content can be bypassed by captions, whether fine-grained visual evidence must be preserved at instance or pixel granularity, or whether visual evidence changes over time. For example, although Mem-Gallery introduces knowledge conflicts, these conflicts are primarily textual rather than visual-state updates. MemEye therefore treats visual evidence as the central memory bottleneck: each item specifies what granularity of visual evidence must survive in memory and how that evidence must be used over time.

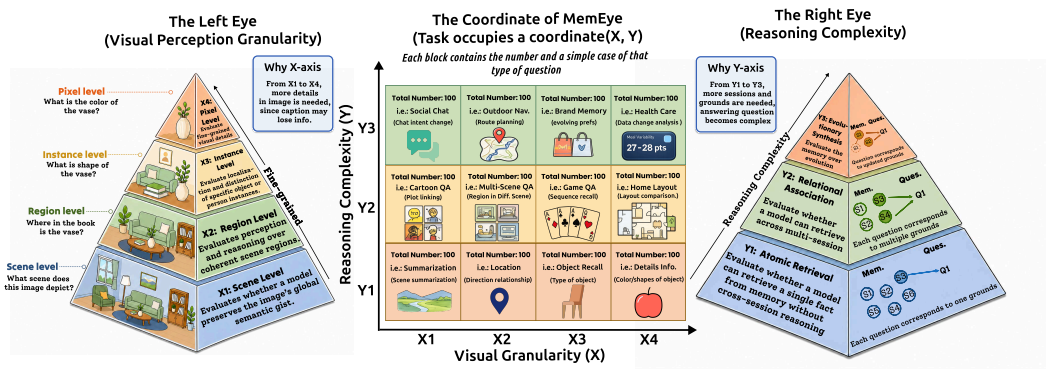


Figure 3: The MemEye two-axis taxonomy. The X-axis captures the granularity of decisive visual evidence, while the Y-axis captures the required reasoning operation over memory.

3 MemEye Framework and Benchmark

3.1 The Two-Dimensional Evaluation Framework

MemEye's evaluation framework is organized as follows. As shown in Figure 3, MemEye contains two dimensions that form a coordinate system. The X-axis represents the first dimension, defined by the granularity of visual perception. From X_1 to X_4, the granularity of the required visual evidence becomes increasingly fine-grained. The definitions of X_1 to X_4 are as follows:

Scene-level. X_1 measures whether the model preserves scene-level evidence, including scene type, activity, and global semantic gist. This level captures the coarsest form of visual evidence and serves as the baseline for visual granularity.
Region-level. X_2 evaluates the ability to perceive and reason over semantically coherent regions. The model must identify meaningful subregions and capture their local context and interactions. Unlike scene-level understanding, this level focuses on localized semantics, where information is organized within regions rather than the entire scene.
Instance-level. X_3 evaluates whether the model can localize and distinguish specific object or person instances within and across images. The key challenge is preserving entity identity when multiple similar candidates exist. Caption-based representations often flatten such distinctions, marking the transition from region-level understanding to instance-level visual memory.
Pixel-level. X_4 requires reasoning over pixel-level evidence, including fine-grained details such as color, texture, or small text. These cues are often absent from text, reflecting pixel-level necessity where critical evidence exists only in the visual signal.

The Y-axis corresponds to the second dimension, capturing the reasoning depth required for memory retrieval during question answering. It emphasizes not only whether sufficient evidence can be located, but also whether that evidence can be associated, revised, and synthesized into the valid answer. This dimension is organized into three levels, from Y_1 to Y_3, as detailed below:

Atomic Retrieval. Y_1 measures whether the model can retrieve a single fact from memory without cross-session reasoning. It primarily tests basic memory access rather than composition, serving as the lowest reasoning baseline.
Relational Association. Y_2 evaluates the ability to associate distributed evidence across sessions and modalities. The reasoning remains monotonic: information accumulates without contradiction. This level captures referential resolution and implicit memory traversal beyond isolated retrieval.
Evolutionary Synthesis. Y_3 tests non-monotonic synthesis over evolving memory. The model must handle updates, conflicts, and overrides, maintaining a coherent world state under revision. Answers are not explicitly stated but must be inferred from changing evidence.

The two dimensions form MemEye's framework. Each question is assigned an (X, Y) coordinate indicating its level of visual evidence and depth of reasoning over memory. The middle sub-figure of Figure 3 illustrates this assignment.

3.2 The Proposed Benchmark

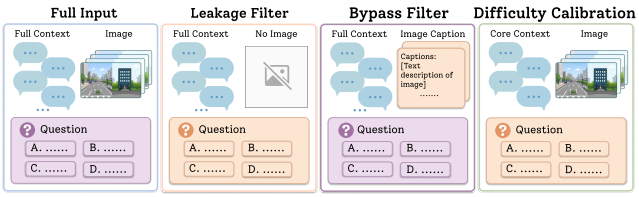


Figure 4: The filtering process used to build the benchmark.

The benchmark contains 1,200 questions across 344 sessions and 940 images, with 100 questions in each of the 12 (X, Y) taxonomy cells. Each question has two mirrored forms (a multiple-choice version and an open-ended version). For MCQ questions, to mitigate VLM bias, we create four rotated variants with the correct answer cycling through A–D. As shown in Figure 1, the benchmark spans eight tasks grouped into four life-scenario domains: *Leisure* (Card Playlog and Cartoon Entertainment), *Domestic* (Home Renovation and Outdoor Navigation), *Professional* (Brand Memory and CrossScene Memory), and *Personal* (Health Care and Social Chat). The images come from both public and archival media, as well as generated content, covering a wide range of image types, including photographs, screenshots, comic panels, and user interface renderings. Each question receives the most demanding X and Y labels needed to answer it; the full (X, Y) cell distribution is provided in Appendix A.2. After generating the candidate visual-memory questions and assigning the label to each question, we use three mechanisms to verify that our benchmark is visual-centric and that these questions arise from limitations in the agent's memory capabilities, rather than from the underlying foundation model.

To be more specific, we perform three filtering mechanisms on each question. To mitigate VLM bias, we use four answer rotations during these checks. **(1) Eliminating answer leakage in dialogue.** For each question, we provide only the question, answer choices, and gold clue-round text, with no images or captions, and test whether the agent can answer correctly across answer rotations. If so, the item is considered solvable without visual evidence and is removed. **(2) Eliminating visual bypassability via minimal captions.** We replace each image with a very short caption and test whether the question can still be answered. The caption only keeps the rough image type, such as a room photo, a game board, or a phone screenshot. If a candidate can still be answered from these captions, we revise or remove it because its visual evidence is too easily replaced by text and therefore does not satisfy the visual-centric requirement. **(3) Controlling for problem difficulty.** We provide the image along with the answer-relevant context to assess whether the question is inherently solvable. This setting does not evaluate memory; rather, it isolates answerability. If the model fails, the difficulty is attributed to limitations of the underlying foundation model rather than its memory. Through these mechanisms, we retain questions that require visual information and are suitable for evaluating memory capabilities rather than only foundation-model recognition ability. More details are provided in Appendix A.4.

4 Experiments and Analysis

In this section, we use our benchmark to analyze current multi-modal agent memory systems. Our analysis moves from locating failures to explaining their causes. We first validate the rationality of our framework's configuration using our benchmark. Then we ask three questions: **RQ1**: Where do current memory systems fail in the MemEye matrix? **RQ2**: Why do memory systems lose visual information? **RQ3**: Why do memory systems lose evolving visual states? Together, these questions first map the failure landscape, then isolate the visual-evidence bottleneck in high-X questions, and finally diagnose why retrieval remains insufficient when memory evidence evolves over time.

4.1 Evaluation Setup

Models and memory methods. We evaluate up to 13 methods across 4 model backbones including Qwen3-VL-8B-Instruct (Bai et al., 2025), GPT-4.1-nano, GPT-5.4-mini (OpenAI, 2026), and Gemini-2.5-flash-lite (Google, 2026). All runs use deterministic decoding.

The evaluated methods include seven text-based memory approaches and six multimodal memory approaches. The text-based methods are Full Context (FC(T)), Semantic RAG (SRAG(T)), Reflexion (Refl.) (Shinn et al., 2023), Generative Agents (Gen.Ag.) (Park et al., 2023), MemoryOS (MemOS) (Kang et al., 2025), A-Mem (Xu et al., 2026), and SimpleMem (SM(T)) (Liu et al., 2025). These methods replace each image with a dense GPT-5.2 caption. The multimodal methods are Full Context (FC(V)), Semantic RAG (SRAG(V)), MIRIX (Wang & Chen, 2025), MMA (Lu et al., 2026),

Table 2: Main results on the 1,200-question MemEye evaluation matrix using `gpt-5.4-mini`. Columns correspond to memory methods, grouped into text-only and multimodal families. Within each coordinate, EM is reported for multiple-choice questions, while LLM-as-a-Judge (LLM-Judge) is reported for free-response questions. Avg. is the unweighted mean across the 12 taxonomy cells. Within each metric row, the highest and second-highest distinct scores are highlighted with orange and blue backgrounds, respectively; ties receive the same highlighting. Results on other backbones are shown in Appendix D.

Y	X	Metric	Textual memory							Multimodal memory					
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MIRIX	MMA	M2A	SM(V)
Y1	X1	EM	0.82	0.76	0.69	0.23	0.72	0.67	0.82	0.73	0.77	0.60	0.56	0.57	0.75
		LLM-Judge	0.60	0.30	0.60	0.10	0.30	0.30	0.55	0.81	0.50	0.34	0.45	0.36	0.71
	X2	EM	0.70	0.68	0.51	0.24	0.51	0.61	0.59	0.78	0.81	0.46	0.43	0.45	0.46
		LLM-Judge	0.25	0.38	0.20	0.00	0.50	0.38	0.35	0.79	0.76	0.45	0.33	0.31	0.56
	X3	EM	0.58	0.50	0.42	0.25	0.49	0.49	0.35	0.58	0.61	0.47	0.69	0.50	0.47
		LLM-Judge	0.46	0.35	0.42	0.08	0.38	0.31	0.16	0.59	0.57	0.37	0.68	0.33	0.40
	X4	EM	0.57	0.58	0.38	0.26	0.49	0.53	0.40	0.58	0.69	0.46	0.69	0.36	0.41
		LLM-Judge	0.40	0.42	0.38	0.13	0.38	0.35	0.25	0.42	0.65	0.32	0.54	0.39	0.41
Y2	X1	EM	0.63	0.57	0.60	0.27	0.52	0.51	0.75	0.67	0.66	0.35	0.57	0.42	0.34
		LLM-Judge	0.46	0.54	0.55	0.23	0.32	0.46	0.35	0.50	0.63	0.42	0.44	0.31	0.23
	X2	EM	0.48	0.45	0.44	0.24	0.55	0.55	0.43	0.69	0.45	0.43	0.41	0.32	0.32
		LLM-Judge	0.33	0.10	0.25	0.10	0.33	0.33	0.11	0.20	0.35	0.33	0.46	0.32	0.12
	X3	EM	0.45	0.56	0.39	0.26	0.48	0.42	0.39	0.58	0.69	0.39	0.40	0.37	0.34
		LLM-Judge	0.48	0.50	0.41	0.30	0.45	0.52	0.30	0.60	0.70	0.31	0.50	0.38	0.40
	X4	EM	0.43	0.37	0.29	0.26	0.39	0.38	0.54	0.45	0.44	0.37	0.41	0.24	0.32
		LLM-Judge	0.51	0.45	0.44	0.34	0.60	0.44	0.37	0.47	0.50	0.21	0.24	0.32	0.35
Y3	X1	EM	0.85	0.81	0.85	0.25	0.78	0.78	0.76	0.82	0.88	0.50	0.53	0.43	0.49
		LLM-Judge	0.63	0.63	0.63	0.50	0.50	0.38	0.73	0.54	0.69	0.47	0.59	0.48	0.47
	X2	EM	0.67	0.69	0.64	0.25	0.67	0.73	0.65	0.68	0.77	0.37	0.60	0.40	0.50
		LLM-Judge	0.48	0.32	0.47	0.11	0.49	0.58	0.32	0.30	0.37	0.28	0.30	0.38	0.31
	X3	EM	0.49	0.41	0.39	0.26	0.59	0.66	0.69	0.46	0.51	0.43	0.56	0.43	0.38
		LLM-Judge	0.35	0.39	0.39	0.50	0.25	0.63	0.57	0.30	0.58	0.33	0.54	0.40	0.48
	X4	EM	0.33	0.34	0.33	0.24	0.33	0.44	0.35	0.43	0.57	0.39	0.33	0.28	0.30
		LLM-Judge	0.23	0.23	0.35	0.15	0.23	0.48	0.11	0.32	0.32	0.22	0.39	0.24	0.20
Avg.	–	EM	0.58	0.56	0.49	0.25	0.54	0.56	0.56	0.62	0.65	0.44	0.52	0.40	0.42
		LLM-Judge	0.43	0.38	0.42	0.21	0.39	0.43	0.35	0.49	0.55	0.34	0.46	0.35	0.39

M2A (Feng et al., 2026), and SimpleMem (SM(V)) (Liu et al., 2026). These methods operate on the original visual inputs. For retrieval-based methods, we use top-K=10 and standardize text and image embedding backbones where possible. The full model identifiers, embedding settings, context budgets, and implementation details are provided in Appendix C.5 and C.1. We follow each method's official or recommended retrieval stack when available, so method comparisons should be interpreted as system-level comparisons rather than encoder-controlled ablations.

Metrics and diagnostics. For multiple-choice evaluation, we report exact-match accuracy (EM) averaged over the four answer rotations. For open-ended evaluation, we use LLM-as-a-Judge as the primary metric. We report BLEU-1 as an auxiliary lexical metric in Appendix D.1. To validate the judge, we conduct a human-judge agreement study on a stratified sample of 72 predictions. The automated accept/reject judgments show strong agreement with human labels, with Cohen's $\kappa = 0.94$. Details are provided in Appendix C.2.

4.2 Validation of MemEye

Before reporting diagnostic findings, we verify that the MemEye axes discriminate as intended: X captures visual evidence granularity, and Y captures reasoning depth over memory.

Caption-Proof Diagnostic. To validate the X-axis, we compare native-image memory with dense-caption memory and measure $\Delta = \text{Acc}_{\text{image}} - \text{Acc}_{\text{caption}}$. During benchmark construction, all X_1–X_4 items already pass a minimal-caption bypass filter, removing questions answerable from very short captions that only preserve coarse image type. Here, GPT-5.2 dense captions serve as a stronger textual substitute for testing how much visual evidence is lost when images are stored as

text. If the X-axis captures visual granularity, the image-caption gap should be smaller for scene- and region-level evidence and larger for instance- and pixel-level evidence. Detailed results are reported in Appendix B.1 and analyzed in §4.4.

Oracle-Evidence Diagnostic. To validate the Y-axis, we evaluate an oracle-evidence setting where each question is answered using its ground-truth rounds and original images, removing retrieval as the main bottleneck. Here, "oracle" means that the annotated gold clue rounds are provided directly, rather than retrieved by the memory system. The results are shown in Appendix Table 8. In this setting, GPT-5.4-mini shows a steady drop in LLM-as-a-Judge performance from Y_1 to Y_3 ($0.673 \rightarrow 0.601 \rightarrow 0.558$), indicating that the Y-axis captures reasoning depth beyond retrieval. System-level results are consistent: retrieval-based methods perform well in Y_1, while full-context or state-aware methods become more competitive in Y_3. Thus, the Y-axis reflects differences in memory usage, not just task difficulty.

4.3 RQ1: Where Do Current Memory Systems Fail in the MemEye Matrix?

Table 2 reports cell-level results under EM and LLM-as-a-Judge. Current systems are far from saturating MemEye. At the aggregate level, SRAG(V) achieves the best open-ended performance with LLM-Judge $= 0.55$ and the best multiple-choice performance with EM $= 0.65$. The gap between EM and LLM-as-a-Judge is informative: multiple-choice accuracy can benefit from answer options and broad context coverage, whereas open-ended evaluation reveals whether the system can articulate the relevant memory state.

The results reveal two interacting stressors rather than a single memory challenge. First, fine-grained visual evidence exposes failures that are not visible at the scene level. At low X, caption-based memory remains competitive; at high X, native visual memory becomes more important. For example, at (X_3, Y_1), SRAG(V) reaches an LLM-as-a-Judge score of 0.57, outperforming the best text-based method, FC(T), which reaches 0.46. At (X_4, Y_1), SRAG(V) reaches an LLM-as-a-Judge score of 0.65, while the best text-based method reaches 0.42.

Second, evolving-state reasoning changes the bottleneck after evidence is retrieved. Retrieval works well when the relevant evidence can be selected directly: SRAG(V) remains competitive at (X_2, Y_2) and leads LLM-as-a-Judge performance at (X_3, Y_2). In Y_3 cells, however, the system must decide which evidence remains valid after updates or conflicts. Here, textual memory methods lead open-ended performance: SimpleMem(T) achieves the highest LLM-as-a-Judge score at (X_1, Y_3), while A-Mem leads at (X_2, Y_3), (X_3, Y_3), and (X_4, Y_3). Still, no method solves both axes at once: textual or agentic memory can help organize evolving states but may lose fine visual details, whereas image-based memory preserves more visual evidence but struggles to select the updated visual state. This motivates RQ2 and RQ3, which separately analyze visual-evidence loss and state-selection failure.

4.4 RQ2: Why Do Memory Systems Lose Visual Information?

We next analyze why fine-grained visual evidence is often lost. Current multimodal memory systems differ in how they represent and retrieve visual evidence. For example, MIRIX extracts structured memories from multimodal inputs, while Omni-SimpleMem combines textual summaries with links to raw content. In our evaluation, text-based memory systems receive dense captions in place of images, while multimodal memory systems such as SimpleMem(V), MMA, and M2A retain or recover native image evidence for answering. To compare these storage schemes, we focus on Y_1, where each question corresponds to a single evidence source and does not require multi-hop reasoning. This setting isolates the agent's ability to understand and preserve visual information. Text-based storage methods perform as well as image-based methods on coarse-grained questions (e.g., X_1 and X_2), while image-based methods excel on fine-grained questions (e.g., X_3 and X_4). We attribute this difference to the nature of the two representations: text can capture high-level, generalized descriptions, whereas native images can better preserve fine-grained visual details.

To quantify this effect, we compare each text-based method with its visual counterpart and compute the Caption-Proof gain, $\Delta = \text{Score}_V - \text{Score}_T$. Figure 5(b) reports the average LLM-as-a-Judge gain across the MemEye matrix, with method-specific heatmaps provided in Appendix B.1. Image-based memory helps most when the decisive evidence is fine-grained. In the average heatmap, gains are small in scene-level regions and become positive in fine-grained cells. Bootstrap confidence intervals,

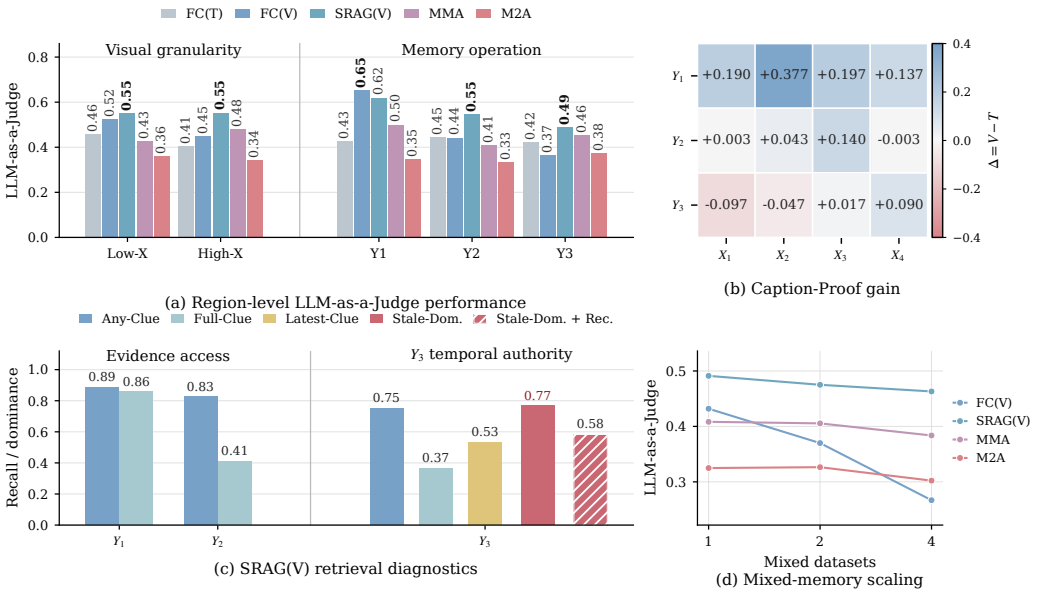


Figure 5: Experimental diagnostics on MemEye under `gpt-5.4-mini`. (a) Region-level LLM-as-a-Judge performance. Bars report coordinate-balanced macro-averages: performance is first averaged within each relevant (X, Y) coordinate, and the resulting coordinate scores are then averaged equally. (b) Average Caption-Proof gain, $\Delta = \text{Score}_V - \text{Score}_T$, shows where native visual evidence improves over dense captions. (c) SRAG(V) retrieval diagnostics separate evidence access from temporal-authority failures; the hatched bar shows a retrieval-only recency counterfactual. (d) Cross-topic dialogue scaling evaluates robustness as unrelated histories from other tasks are added. Results are averaged across two controlled four-task combinations. Full method-specific Caption-Proof heatmaps are reported in Appendix B.1. Together, these panels show current systems trade off visual evidence preservation, evidence selection, and robustness, rather than solving all MemEye regions uniformly.

reported in Table 10, are consistent with this diagnostic pattern. Overall, these results suggest that caption-based storage is more likely to lose decisive instance- and pixel-level evidence. Moreover, Appendix Table 8 shows that, when the correct clue rounds are provided, the gap between text-based and multimodal methods widens as the required visual evidence becomes more fine-grained. More results and analysis are provided in Appendix D.2.

4.5 RQ3: Why Do Memory Systems Lose Evolving Visual States?

RQ2 shows that native image evidence improves fine-grained visual preservation, especially in high-X regions. However, this benefit weakens in Y_3 (Figure 5(b)), where the answer depends on which visual state remains valid after later updates. These cases are not static visual recall: the image-grounded evidence itself changes across sessions. This exposes a second bottleneck beyond visual preservation, which we call **evolving visual state tracking**.

To isolate evolving-state tracking, we compare representative memory systems with oracle evidence controls on the evolving visual-state subset. Table 15 shows that directly providing only the latest valid visual state yields performance close to the full oracle evidence chain. This near match suggests that, for many Y_3 questions, answering correctly depends primarily on identifying the currently valid visual state. In contrast, memory systems remain far below these oracle settings, indicating that they fail not because the final state is visually unreadable, but because they do not recover and prioritize it from an evolving memory history.

Retrieval diagnostics show that related evidence is not enough. As shown in Figure 5(c), Appendix D.4, and the case studies in Appendix Figure 12, SRAG(V) often retrieves evidence about the right topic in Y_3 questions, but can miss the decisive latest clue or the complete update chain. In other words, semantic similarity does not guarantee temporal validity. Table E.1 further illustrates this failure mode: text-based memory methods can preserve compact evolving-state evidence chains,

while multimodal memory methods are distracted by visually similar, stale, or conflicting retrieved images. In other words, finding related evidence is not the same as selecting the valid state. If retrieval fails to select the updated evidence, one possible alternative is full-context conditioning: provide the entire history and let the model resolve the valid state. We further conduct a memory scaling analysis by concatenating conversations from different tasks into a larger, more diverse memory history. The results in Figure 5(d) show that memory mechanisms become increasingly important as history length and topic diversity grow, because they help filter unrelated evidence. These findings suggest that multimodal memory should combine image evidence, text, or structured state records, and mechanisms for selecting valid evidence over long histories. Together, these results show that Y3 failures arise from incomplete evolving-state tracking: systems must not only preserve visual details, but also recover the update chain and prioritize the currently valid visual state over stale evidence.

5 Implications for Memory Architecture Design

The results above suggest that current multimodal memory should not be designed as a single retrieval module. MemEye's two axes point to complementary design requirements: the X-axis calls for preserving decisive visual evidence at the right granularity, while the Y-axis calls for selecting the temporally valid evidence once memory evolves.

How to store visual information? RQ2 shows that agent memory systems lose visual information when images are converted into text. Text-based memory can capture coarse visual information, but it often misses fine-grained visual information. Image-based memory is therefore needed for high-X questions, where the answer depends on visual information that captions may omit. Thus, future memory systems should preserve image evidence rather than relying only on text summaries.

How to select valid visual memory states? RQ3 shows that preserving image evidence is still not enough. When state information changes across sessions, the system must decide which visual state is currently valid. Text-based or structured memory can do better for high-Y questions in recording updates, conflicts, and overrides, while image-based memory preserves the visual information needed to check those states. The cross-topic dialog scaling ablation further shows why memory mechanisms are useful as memory grows: full-context methods become more sensitive to unrelated histories, while retrieval-based or structured memory methods remain more stable. This suggests that memory systems need mechanisms that filter, compress, or reweight evidence before answering. At the same time, current mechanisms are not yet state-aware enough: retrieval diagnostics show that semantically relevant evidence can still be stale. The recency re-ranking probe in Appendix D.4 shows that temporal signals reduce stale-over-latest errors, but do not fully solve answer generation.

Together, these findings suggest that multimodal memory should combine image-based and text-based representations. Image-based memory preserves fine-grained visual information, while text-based or structured memory helps track state changes across sessions. On top of both, the system needs mechanisms for selecting valid evidence by filtering unrelated history, using temporal signals, and avoiding stale evidence. A promising direction is to combine image evidence, structured state records, and recency-aware selection within a unified memory system. We discuss benchmark limitations and directions for broader evaluation in Appendix D.7.

6 Conclusion

We introduced MemEye, a visual-centric benchmark for evaluating long-term multimodal memory. Its two-axis design separates visual evidence granularity from memory reasoning depth, covering settings from coarse scene-level information to fine-grained visual details and from single-evidence retrieval to reasoning over evolving states. Based on this framework, we construct 1,200 questions across eight life-scenario tasks, with balanced coverage of the 12 taxonomy cells and validation procedures for answerability, shortcut resistance, visual dependence, and reasoning structure. Experiments with 13 memory methods and four VLM backbones show that current systems still struggle to preserve fine-grained visual information, retrieve temporally valid evidence, and track which visual state remains current after updates. These results point to the need for memory systems that preserve visual detail while also supporting reliable state tracking over long interactions.

AI Use Statement

In this paper, we used AI tools solely to refine grammar and improve the clarity and readability of the text.

Ethics Statement

All datasets used in this work are publicly available and contain no sensitive information. Our method enhances LLM reasoning without introducing new data collection or human interaction. While stronger reasoning ability may be misused, we emphasize that this work is intended for beneficial research and responsible applications.

Reproducibility Statement

We will release our code and data once the paper is published. The appendix includes detailed experimental setups and hyperparameters so that others can reproduce our results. We also encourage the community to follow good research practices when using our code and data, to help maintain the reliability and transparency of future work.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105 (38):14325–14329, 2008.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. Twinvoice: A multi-dimensional benchmark towards digital twins via llm persona simulation. *arXiv preprint arXiv:2510.25536*, 2025.

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. *arXiv preprint arXiv:2602.07624*, 2026.

Google. Gemini api model documentation. https://ai.google.dev/gemini-api/docs/models, 2026. Accessed: 2026-05-01.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=DT7JyQC3MR.

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

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)*, pp. 1100–1110, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.123. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.123.

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*, pp. 25961–25970, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 582–595, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1058. URL https://aclanthology.org/N19-1058/.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, Bowon Ko, and Ho-Jin Choi. Multiverse: A multi-turn conversation benchmark for evaluating large vision and language models. *arXiv preprint arXiv:2510.16641*, 2025.

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*, pp. 29735–29762, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1511. URL https://aclanthology.org/2025.emnlp-main.1511/.

Jiaqi Liu, Yaofeng Su, Peng Xia, Yiyang Zhou, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2025. URL https://github.com/aiming-lab/SimpleMem.

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. *arXiv preprint arXiv:2604.01007*, 2026.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. Convbench: A multi-turn conversation evaluation

benchmark with hierarchical ablation capability for large vision-language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=PyTf2jj0SH.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=s8h2jSN6a6.

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to me: Long-term personalized referential memory qa. *arXiv preprint arXiv:2603.01990*, 2026. doi: 10.48550/arXiv.2603.01990. URL https://arxiv.org/abs/2603.01990.

OpenAI. Openai api model documentation. https://platform.openai.com/docs/models, 2026. Accessed: 2026-05-01.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

Zeru Shi, Kai Mei, Mingyu Jin, Yongye Su, Chaoji Zuo, Wenyue Hua, Wujiang Xu, Yujie Ren, Zirui Liu, Mengnan Du, Dong Deng, and Yongfeng Zhang. From commands to prompts: LLM-based semantic file system for AIOS. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2G021ZqUEZ.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Turing Motors. Japan open driving dataset sample. https://huggingface.co/datasets/turing-motors/Japan-Open-Driving-Dataset-Sample, 2024. Accessed: 2026-05-01.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents, 2025. URL https://arxiv.org/abs/2507.07957.

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

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FiM0M8gcct.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, and Yu Qiao. MMRC: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22477–22503, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1096. URL https://aclanthology.org/2025.acl-long.1096/.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1850–1860, 2025.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6), September 2025. ISSN 1046-8188. doi: 10.1145/3748302. URL https://doi.org/10.1145/3748302.

A Benchmark Construction and Dataset Details

A.1 Task Statistics

Table 3 provides per-task statistics for the MemEye benchmark. The main results in Table 2 and the corresponding performance and visual-gain heatmaps use the full 1200-question benchmark. Additional experiments reported in this appendix use subsets of questions (371 questions).

Table 3: Per-task statistics of the MemEye benchmark.

Task	Sessions	Rounds	Images	Questions
Home Renovation	26	224	180	238
Brand Memory	42	72	30	48
Card Playlog	4	30	30	71
Cartoon Ent.	118	331	119	129
CrossScene Memory	54	363	225	262
Outdoor Nav.	34	204	136	155
Health Care	27	230	177	231
Social Chat	39	53	43	66
Total	344	1507	940	1200

Image provenance. Table 4 marks each task as archival/public (**A**) or generated/rendered (**G**). Brand Memory uses advertisement images from the Pitt Image Ads dataset (Hussain et al., 2017). Cartoon Ent. uses scanned pages from the public-domain comic strip *Alley Oop* and story illustrations from Seed-Story (Yang et al., 2025). Home Renovation uses stock interior-design photographs. Card Playlog uses HTML-rendered game-state screenshots based on Cardiverse (Li et al., 2025). Social Chat uses synthetic face images from StyleGAN (Karras et al., 2019) (no real people) combined with PIL-rendered chat-UI screenshots. Outdoor Nav. uses dashcam frames from the Japan Open Driving Dataset (Turing Motors, 2024). CrossScene Memory and Health Care use AI-generated images produced with `gpt-5.2` (DALL-E) under controlled scene blueprints.

A.2 Taxonomy Cell Distribution

Table 5 shows the number of questions in each (X_i, Y_j) cell of the two-axis taxonomy.

A.3 Detailed Taxonomy Definitions

This section provides the detailed annotation definitions behind the compact taxonomy description in §3. MemEye assigns each question an (X, Y) coordinate using a highest-bottleneck rule: the X

Table 4: MemEye dataset composition across eight life-scenario tasks. Each task contains n multiple-choice questions and an equal number of mirrored open-ended questions. Source denotes whether visual evidence is archival/public (**A**) or generated/rendered (**G**).

Domain	**Task**	**Scenario type**	**Src.**	**Primary** (X, Y)	n
Leisure	Card Playlog	game-state tracking	G	X_4, Y_2–Y_3	71
Leisure	Cartoon Ent.	character and narrative memory	A	X_1–X_2, Y_1–Y_3	129
Domestic	Home Renovation	room-state and design updates	A	X_3–X_4, Y_2–Y_3	238
Domestic	Outdoor Nav.	route and landmark memory	G	X_3, Y_1–Y_2	155
Professional	Brand Memory	logo and visual identity recall	A	X_2–X_4, Y_1–Y_2	48
Professional	CrossScene Memory	object-state updates across scenes	G	X_2–X_4, Y_2–Y_3	262
Personal	Health Care	dashboard and portal updates	G	X_1–X_4, Y_1–Y_3	231
Personal	Social Chat	personal visual-detail memory	G	X_2–X_3, Y_1–Y_2	66
Total: 8 tasks across 4 life-scenario domains					**1,200**

Table 5: Question distribution across the (X, Y) taxonomy matrix.

	X_1	X_2	X_3	X_4	Total
Y_1	100	100	100	100	400
Y_2	100	100	100	100	400
Y_3	100	100	100	100	400
Total	300	300	300	300	1,200

label is determined by the finest level of decisive visual evidence required for the answer, and the Y label is determined by the deepest memory operation required after the relevant evidence is available.

Table 6 summarizes the definitions used to assign axis labels and presents representative MemEye examples for each axis level. The examples are intentionally drawn from different task domains to show that the taxonomy is attached to evidence requirements and memory operations rather than to task names.

A.4 Item-Level Filtering and Taxonomy Audit Details

Stage 2 filters and audits candidate questions before they enter the locked benchmark. All MCQ-based gates use four-way answer rotations and a two-level model panel consisting of `gpt-5.4-mini` and `gpt-5.2`. For shortcut-resistance gates, robust success means that both panel models answer all four rotations correctly (rotation-averaged EM $= 1.0$). For the oracle answerability gate, a candidate is treated as unsolved if both panel models remain at or below chance level under oracle visual evidence, i.e., rotation-averaged EM ≤ 0.25. A candidate is flagged when shortcut success is robust across rotations, when it remains unsolved under oracle visual evidence, or when its clue structure does not match the intended Y label. Flagged candidates are revised, removed, or relabeled.

Shortcut-Resistance Gates. We first test whether a candidate can be answered without the intended multimodal memory signal. In the option-only gate, the model receives only the question and rotated answer options, with no context, image, or caption. Robust success indicates that the option set, world knowledge, or stylistic artifacts leak the answer, so the item is flagged. In the text-only clue gate, the model receives the question, options, and clue-round text, but no image or caption. Robust success indicates textual leakage from the dialogue, so the item is flagged. In the minimal-caption gate, original images are replaced by very short captions that only keep the rough image type, such as a room photo, a game board, a cartoon panel, or a phone screenshot. This gate is applied to all X_1–X_4 candidates. Robust success indicates that the question is too easily answered without the original image, so the item is flagged for revision or removal. This gate does not define the X label; the X label is assigned by the finest visual evidence needed to answer the question.

Table 6: Detailed MemEye taxonomy definitions with representative dataset examples.

Level	Definition	Representative MemEye example
Visual evidence granularity (X)		
X_1 Scene-level	Global scene type, activity, or semantic gist; often recoverable from a good caption.	In Cartoon Ent., identifying the open blue-green sky and clouds behind a baby dinosaur climbing a tree.
X_2 Region-level	Semantically coherent subregions, grouped entities, or local spatial structure within a scene.	In Home Renovation, deciding which cabinet sample is closest to a brass hardware piece lying on the floor.
X_3 Instance-level	Specific object or person identity among visually or semantically similar candidates.	In Home Renovation, matching the labeled cabinet sample from a three-sample comparison to the same sample later shown beside a tape measure and pencil.
X_4 Pixel-level	Fine visual details such as small text, exact color, texture, count, or OCR-like evidence.	In CrossScene Memory, reading the current identification tag number displayed inside the fossil-room case.
Memory-reasoning depth (Y)		
Y_1 Atomic retrieval	One sufficient evidence unit answers the question; no cross-session composition is needed.	A single clue image in Cartoon Ent. is enough to answer what the background looks like in the tree-climbing scene.
Y_2 Relational association	Multiple non-conflicting clues must be linked across sessions, modalities, or references.	In Cartoon Ent., a palace-arc question requires comparing two remembered events to decide whether the crowd pelting or dinner fight occurred first.
Y_3 Evolutionary synthesis	Temporally ordered clues include updates, conflicts, or overrides; the answer depends on the valid current state.	In the fossil-room case, an earlier identification tag is later replaced, so the answer must use the updated tag rather than the stale one.

Oracle Answerability Gate. We then provide the gold-clue rounds and the original images. This oracle-evidence setting removes memory search as the bottleneck and checks whether the candidate is well-defined, visually answerable, and within the intended difficulty range. Candidates that remain unsolved under oracle visual evidence are treated as visually ambiguous, underspecified, or above the intended difficulty level, and are flagged. This gate ensures that MemEye evaluates memory access and evidence use rather than impossible visual recognition.

Taxonomy-Structure Audit. Finally, we audit whether the clue structure matches the annotated Y level. A Y_1 item must be answerable from one evidence unit. A Y_2 item must require associating multiple non-redundant evidence units under monotonic accumulation; no single clue should collapse it into atomic retrieval. A Y_3 item must require temporally ordered evidence in which a later clue updates, overrides, or conflicts with an earlier state; the answer must depend on resolving the valid current state rather than retrieving either clue in isolation. This audit ensures that the Y-axis reflects evidence use rather than task naming.

A.5 Annotation Agreement

MemEye questions are initially generated by `gpt-5.2` conditioned on target (X, Y) coordinates, then reviewed and adjudicated by human annotators who inspect the visual evidence, clue rounds, and question–answer pairs. The agreement study below is a reproducibility check on the taxonomy definitions, not a substitute for human adjudication. To quantify label reproducibility, we measure inter-annotator agreement on a stratified subset of 100 questions using two strong LLMs as independent annotators: GPT-5.4 and Gemini-2.5-Pro.

Protocol. We sample 100 questions stratified across all (X, Y) cells (8–10 per cell). Each model receives the question, ground-truth answer, annotator explanation, and the full taxonomy definitions,

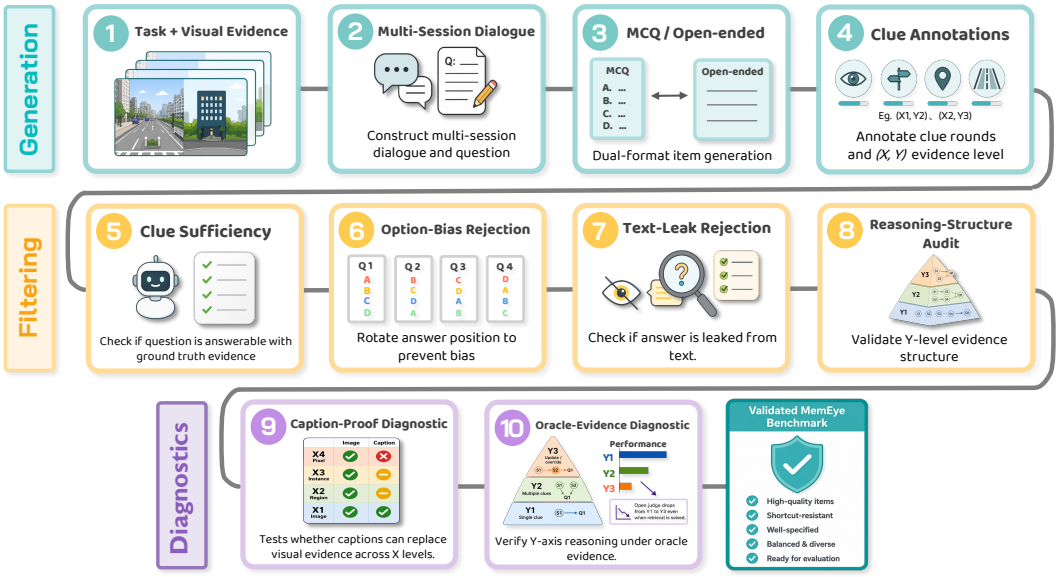


Figure 6: Complete MemEye construction and validation pipeline. Candidate questions are generated from task scenarios, visual evidence, and target (X, Y) regions, then authored as mirrored MCQ/open-ended items with clue-round annotations and four-way MCQ rotations. Item-level checks establish multimodal validity, memory answerability, and taxonomy alignment by removing shortcut-answerable or under-specified candidates and verifying that the clue structure matches the annotated Y level. Finally, aggregate diagnostics validate the two benchmark axes: caption substitution probes X-axis visual irreplaceability, and oracle evidence probes Y-axis reasoning over retrieved memory.

including boundary examples and decision rules (e.g., "X2 is about WHERE things are in a local region; X3 is about WHICH specific instance is being referred to among similar candidates"). The models independently assign X and Y labels. Neither model has access to the original images; labels are assigned from the textual description of visual evidence in the question and explanation. Therefore, this agreement study measures the reproducibility of the written taxonomy definitions and annotation rationales, rather than serving as primary evidence that visual bottlenecks are objectively labeled from images. Final label adjudication is performed by human annotators who inspect the original images, dialogue context, clue rounds, and question-answer pairs. Visual bottleneck validity is evaluated separately through the Caption-Proof diagnostic and caption-robustness ablation.

Results. Inter-annotator agreement is substantial on both axes: $\kappa_X = 0.66$ and $\kappa_Y = 0.63$ (Cohen's κ). The two models agree on 76 of 100 X labels and 77 of 100 Y labels. Remaining X-axis disagreement concentrates on the X_2/X_3 boundary (region-level vs. instance-level), which often requires inspecting the actual image to determine whether the bottleneck is spatial layout or identity binding. Y-axis disagreement is distributed more evenly, with occasional confusion between Y_1 (atomic retrieval) and Y_2 (relational association) when the number of required evidence units is ambiguous from the textual description alone.

Human adjudication. All 1,200 questions in the final benchmark have been reviewed by a human annotator who inspects the images, dialogue context, and clue structure. The LLM-based agreement measurement reported here serves as a reproducibility check on the taxonomy definitions rather than as a replacement for human review. Cases where the two LLM annotators disagreed were cross-checked against the human-adjudicated gold labels; the gold labels were revised only when the disagreement revealed a genuine annotation error rather than a boundary judgment call.

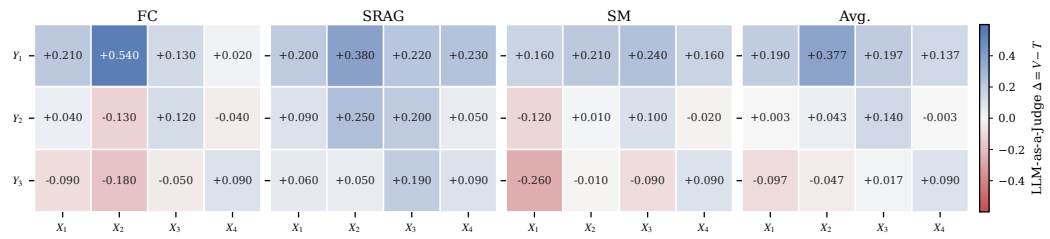


Figure 7: Full Caption-Proof heatmaps on MemEye under `gpt-5.4-mini`. Each cell reports $\Delta = \text{Score}_V - \text{Score}_T$ for matched textual and visual streams, where textual streams use dense captions and visual streams use native images. The main text reports the average heatmap in Figure 5(b).

Table 7: Caption-proof validation on `gpt-5.4-mini`. Rows aggregate over Y_1-Y_3 for each visual granularity level. Judge(T) and Judge(V) denote LLM-Judge scores for matched caption-based and native-visual streams; Δ is the visual-stream gain.

Method	X	n	Judge(T)	Judge(V)	Δ_{J}	Δ_{EM}
FC	X1	49	0.5306	0.5612	+0.0306	+0.0765
FC	X2	42	0.2976	0.2857	-0.0119	-0.0952
FC	X3	144	0.3854	0.4653	+0.0799	+0.1076
FC	X4	136	0.3640	0.4007	+0.0368	+0.0478
SRAG	X1	49	0.5306	0.6020	+0.0714	-0.0510
SRAG	X2	42	0.2738	0.4286	+0.1548	+0.0357
SRAG	X3	144	0.3194	0.6076	+0.2882	+0.1944
SRAG	X4	136	0.3162	0.3493	+0.0331	+0.0772
SM	X1	49	0.5204	0.4694	-0.0510	+0.0612
SM	X2	42	0.2143	0.1667	-0.0476	+0.0238
SM	X3	144	0.2361	0.2812	+0.0451	+0.0260
SM	X4	136	0.2353	0.2132	-0.0221	-0.0110

B Benchmark Validation

B.1 Caption-Proof Validation Details

Caption-Proof validation assesses whether a benchmark question requires native visual evidence, rather than being solvable from captions alone. For each question, we evaluate matched textual and multimodal memory systems under the same memory architecture. The textual memory replaces every image with a dense caption. The multimodal memory keeps the original images. We then report the visual gain $\Delta = \text{score}_V - \text{score}_T$, where positive values indicate that native images preserve critical evidence that captions do not.

Table 7 and Figure 7 report the aggregate Caption-Proof gaps on `gpt-5.4-mini` for the three matched method families used in the main analysis: Full Context, Semantic RAG, and SimpleMem. Rows aggregate over Y_1-Y_3 within each visual-granularity level. The table is intended as an X-axis diagnostic: scene- and region-level evidence are often included in captions, whereas instance identity, exact attributes, small text, and temporally valid visual states are more likely to be lost when images are converted to text. The qualitative cases in Figures 10 and 11 illustrate this failure mode directly.

B.2 Oracle-Evidence Validation Details

Oracle-evidence validation tests whether the annotated axes reflect the intended memory requirements after retrieval difficulty is removed. For each question, the model is given the gold-clue rounds and the original images, while controlling for search over the full history and preserving the evidence needed to answer.

Table 8: Oracle-evidence diagnostics on `gpt-5.4-mini`. Left: performance by memory-reasoning level when gold clue rounds and original images are provided. Right: LLM-Judge visual gap by visual granularity under gold evidence.

(a) Reasoning depth					**(b) Visual granularity**				
Y	n	EM	B-1	Judge	X	n	Text	Visual	Δ
Y_1 Atomic	113	0.856	0.412	0.673	X_1 Scene	49	0.653	0.776	+0.122
Y_2 Rel.	198	0.633	0.426	0.601	X_2 Region	42	0.262	0.524	+0.262
Y_3 Evol.	60	0.696	0.327	0.558	X_3 Inst.	144	0.358	0.622	+0.264
					X_4 Pixel	136	0.335	0.632	+0.298

Table 9: Human oracle sanity check. Three annotators independently answer a stratified subsample of questions given only the gold clue rounds and original images. We report majority-vote accuracy and mean individual accuracy by Y level. This serves as a sanity check rather than a full human ceiling estimate.

Metric	Y_1	Y_2	Y_3	All
Majority-vote accuracy	1.00	0.83	0.81	0.88
Mean individual accuracy	0.97	0.81	0.77	0.85

Table 8(a) validates the Y-axis on `gpt-5.4-mini`. We use open-ended LLM-Judge as the primary metric because multiple-choice options can mask reasoning failures; EM and BLEU-1 are reported as auxiliary metrics. LLM-Judge decreases from Y_1 to Y_3 ($0.673 \rightarrow 0.601 \rightarrow 0.558$), showing that reasoning remains harder even when the relevant evidence is provided. This supports the interpretation of Y_1 as atomic retrieval, Y_2 as relational association, and Y_3 as evolutionary synthesis over updates, overrides, or conflicts.

Table 8(b) validates the X-axis by comparing text-only and multimodal gold-evidence settings. The visual gap increases from $+0.122$ at X_1 to $+0.298$ at X_4, indicating that native visual evidence becomes more important as the decisive evidence shifts from scene-level content to instance- and pixel-level details.

B.3 Human Validation Under Oracle Evidence

To verify that MemEye questions are answerable by humans when the relevant evidence is provided, three annotators independently answer a stratified subsample of MCQ questions given only the gold clue rounds and original images. Annotators do not see the full conversation history. For Y_3 items, clue rounds are presented in chronological order with instructions to answer based on the latest valid state.

Table 9 reports accuracy by Y level. We treat this study as a small-scale human oracle sanity check rather than a full human ceiling estimate. Human annotators achieve near-perfect accuracy on Y_1, while accuracy decreases on Y_3. This lower Y_3 score does not by itself indicate that the questions are visually unanswerable: Y_3 items require resolving updates, conflicts, or overrides across clue rounds, and annotators can still make state-resolution errors even when retrieval is removed. At the same time, the result shows that Y_3 remains harder under oracle evidence, so we do not use this study to claim near-perfect human performance on all evolutionary-synthesis items. Instead, it serves as a sanity check on the sampled questions, while the benchmark construction relies on separate human adjudication and oracle answerability gates to remove ambiguous or under-specified candidates.

C Evaluation Protocol and Implementation

C.1 Benchmark Runner and Method Settings

All backbone runs use deterministic decoding for answer generation, with temperature $T{=}0$. The main `gpt-5.4-mini` configuration uses a maximum generation length of 128 tokens for benchmark answers. For open-ended questions, LLM-as-a-Judge scoring is enabled with `gpt-5.2`; the judge

prompt returns a normalized semantic-correctness score in $[0, 1]$ and a short rationale. Multiple-choice scores are computed from the extracted answer letter and averaged over the four answer-rotation variants of each original question.

We evaluate 4 open- and closed-source VLM backbones: Qwen3-VL-8B-Instruct (Bai et al., 2025), `gpt-4.1-nano` and `gpt-5.4-mini` (OpenAI, 2026), and `gemini-2.5-flash-lite` (Google, 2026). The thirteen memory methods are Full Context (FC(T/V)), Semantic RAG (SRAG(T/V)), A-Mem, MemoryOS, Reflexion (Shinn et al., 2023), Generative Agents (Park et al., 2023), SimpleMem (T/V) (Liu et al., 2025; 2026), MIRIX (Wang & Chen, 2025), MMA (Lu et al., 2026), and M2A (Feng et al., 2026). We evaluate all 13 methods with `GPT-5.4-mini` and `GPT-4.1-nano` and 12 methods with `gemini-2.5-flash-lite` and Qwen3-VL-8B-Instruct.

Textual memory methods operate on a captioned stream: each image-bearing round is converted once into a dense caption using `gpt-5.2`, and the memory method receives the original dialogue text plus these captions rather than the native image. Native multimodal methods receive the original visual inputs. Full-context methods receive the available history directly; retrieval-based methods retrieve a bounded set of memory rounds; agentic and structured memory methods use their method-specific memory-writing and querying procedures under the same benchmark runner.

For retrieval-based comparisons, we set top-K=10 unless a method's official interface requires an internal iteration budget. Text retrieval uses `all-MiniLM-L6-v2`. Semantic RAG (V) and M2A use `siglip2-base-patch16-384` for image embeddings, while MMA uses `google/siglip-so400m-patch14-384` to match its official implementation. For Semantic RAG (V), text and image dense similarities are combined with equal weight. M2A uses its official memory-manager and query-loop budgets, with up to 15 memory-manager iterations and up to 5 query iterations. MMA uses its default confidence-scoring weights for source, time, and consensus evidence. These settings keep the retrieval budget comparable while preserving method-specific mechanisms that are central to each memory architecture.

Embedding backbones. We use each retrieval-based method with the embedding backbone recommended by its original implementation when available. This choice reflects a full-system comparison: the retrieval encoder is part of the deployed memory method rather than a free interchangeable component. Forcing all methods to share a single embedding model can disadvantage methods whose memory writing, indexing, or confidence scoring is designed around a specific encoder. We therefore standardize embeddings where this is compatible with the method design, while preserving official backbones when they are integral to the method implementation. As a result, architecture-level comparisons should be interpreted as comparisons between full memory systems, not isolated encoder ablations.

C.2 LLM-as-a-Judge Validation

To validate the LLM-as-a-Judge metric used for open-ended evaluation, we conduct a human-judge agreement study. We stratify-sample 3 questions per (X, Y) cell and collect the open-ended predictions from two representative methods, FC(V) and SRAG(V), under `gpt-5.4-mini`. This produces 72 prediction-gold-answer pairs. The items are randomized and method labels are hidden to prevent annotator bias. One human annotator independently scores each prediction as Accept (1) or Reject (0), using the same correctness criterion as the automated judge: whether the prediction conveys the same information as the gold answer. We then compare the human labels with the binarized GPT-5.2 judge scores ($\geq 0.5 \rightarrow$ Accept, $< 0.5 \rightarrow$ Reject).

Agreement is 97.2% (69/71 items, excluding 1 borderline case), with Cohen's $\kappa = 0.94$. Only two items produce disagreement, both involving the same navigation question where the gold answer describes a route by perceptual attributes while the model describes it by landmarks. The confusion matrix is nearly diagonal (Accept–Accept: 40; Reject-Reject: 29; off-diagonal: 1 each direction), and per-method agreement is balanced (FC(V): 97.2%; SRAG(V): 97.1%). These results indicate that the automated LLM-as-a-Judge scores closely track human accept/reject judgments on MemEye open-ended predictions.

Table 10: Key paired bootstrap confidence intervals on `gpt-5.4-mini`. Intervals are percentile 95% confidence intervals over 10,000 question-level resamples with seed 20260430. Positive values indicate gains for the first term in each comparison. CI-L and CI-U denote the lower and upper confidence bounds.

Comparison	Slice	n	Mean	CI-L	CI-U
Cap.-Proof Judge	Low-X V–T	91	+0.024	-0.035	+0.081
Cap.-Proof Judge	High-X V–T	280	+0.079	+0.042	+0.115
Cap.-Proof EM	Low-X V–T	91	+0.010	-0.037	+0.058
Cap.-Proof EM	High-X V–T	280	+0.075	+0.043	+0.107
SRAG(V)–FC(V) Judge	All	371	+0.058	+0.005	+0.111
SRAG(V)–M2A Judge	$Y_3 - Y_1$	113/60	-0.425	-0.586	-0.265

C.3 Prompt Examples

This section reports the prompts used in our experiments. We group them by role: answer generation, caption generation, question generation, taxonomy annotation, and LLM-as-a-Judge evaluation. For LLM-as-a-Judge scoring, we follow the prompt-based evaluation protocol used by Mem-Gallery (Bei et al., 2026).

Multimodal Open-Ended Answering

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

Rules: ground every answer in the conversation history and images; do not invent facts; when evidence conflicts, prefer the most recent visual evidence unless the question asks about the conflict; inspect images carefully; be precise about spatial positions, colors, textures, and visual attributes; say the information is absent if it is genuinely unavailable; give only the answer, with no reasoning or restatement. For counting questions, reply with only the number. For yes/no questions, start with "Yes" or "No" and add only essential corrective detail. For descriptive questions, answer in one or two short sentences.

Multimodal Multiple-Choice Answering

You are an AI assistant being evaluated on MemEye, a vision-centric multimodal memory benchmark. Answer multiple-choice questions about a long-horizon multimodal conversation using the provided conversation history and images.

Rules: ground every answer in the conversation history and images; do not invent facts; when evidence conflicts, prefer the most recent visual evidence unless the question asks about the conflict; inspect images carefully; reply with only the option letter, e.g., A, B, C, or D.

Text-Plus-Caption Open-Ended Answering

You are an AI assistant being evaluated on MemEye, a text-plus-caption memory benchmark setting. Answer questions about a long-horizon conversation using the provided conversation history and image captions only.

Rules: ground every answer in the provided conversation history and caption text; do not invent facts; do not assume access to raw image pixels; treat captions as the only visual evidence; if a required visual detail is not stated in the dialogue or captions, say the information is unavailable rather than guessing; be precise and concise; give only the answer. For counting questions, reply with only the number. For yes/no questions, start with "Yes" or "No" and add only essential corrective detail. For descriptive questions, answer in one or two short sentences.

Text-Plus-Caption Multiple-Choice Answering

You are an AI assistant being evaluated on MemEye, a text-plus-caption memory benchmark setting. Answer multiple-choice questions about a long-horizon conversation using the provided conversation history and image captions only.

Rules: ground every answer in the provided conversation history and caption text; do not invent facts; do not assume access to raw image pixels; if a visual detail is not stated in the dialogue or captions, choose only if the answer is supported by the provided text, otherwise make the best grounded choice from the evidence; reply with only the option letter, e.g., A, B, C, or D.

Caption Generation

Generate a detailed caption for this image. 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.

Question Generation

You are helping construct MemEye, a vision-centric benchmark for long-term multimodal agent memory. Generate candidate memory questions for the specified target taxonomy cell.

Inputs. Target visual-evidence level: `{{target_X}}`. Target memory-reasoning level: `{{target_Y}}`. Task scenario: `{{task_scenario}}`. Candidate dialogue rounds, timestamps, images, and available image descriptions: `{{evidence_context}}`.

Taxonomy. X_1 requires scene-level visual evidence; X_2 requires region-level spatial or local-area evidence; X_3 requires identifying a specific object or person instance; X_4 requires fine visual evidence such as exact color, small text, texture, count, or markings. Y_1 requires one sufficient evidence unit; Y_2 requires linking multiple non-conflicting clues; Y_3 requires resolving updates, conflicts, replacements, disappearance/reappearance, or the latest valid state.

Requirements. The question must require the target X and Y levels. It should be answerable from the provided evidence, but should not be answerable from general knowledge, answer-option priors, or dialogue text alone when the target requires visual information. For Y_3, include at least one stale or superseded clue and one later clue that establishes the valid state. Avoid ambiguous wording, subjective judgments, and questions whose answer depends on information not visible or not stated in the evidence.

Output format. Return a JSON object with fields `open_question`, `ground_truth`, `mcq_question`, `options`, `correct_option`, `clue_rounds`, `target_X`, `target_Y`, and `explanation`. The explanation should state the decisive visual evidence, why the item fits the target X level, and why the clue structure fits the target Y level.

Taxonomy Agreement Annotation

You are an independent annotator for MemEye, a benchmark that labels each memory question by visual evidence granularity X and memory reasoning depth Y. Assign one X label and one Y label to the item below.

Inputs. Question: `{{question}}`. Ground-truth answer: `{{ground_truth}}`. Annotator explanation: `{{explanation}}`.

Important instruction. You do not have access to the original images. Use only the question, ground-truth answer, and annotator explanation. If the explanation describes visual evidence, treat it as the available description of what the human annotator saw. Do not solve the question; label the type of evidence and reasoning required to answer it.

Highest-bottleneck rule. Choose the finest X level required by the decisive evidence, and the deepest Y operation required after the relevant evidence is available.

X-axis labels. X_1 Scene-level: the answer depends on global scene gist, activity, or overall context. X_2 Region-level: the answer requires a local area, spatial relation, or grouped region, but not identity of one specific instance among similar candidates. X_3 Instance-level: the answer requires identifying a specific object or person instance, especially among similar candidates or across images. X_4 Pixel-level: the answer depends on fine visual details such as exact color, small text, texture, count, symbol, or OCR-like evidence.

Y-axis labels. Y_1 Atomic Retrieval: one evidence unit is sufficient once found. Y_2 Relational Association: multiple non-conflicting clues must be linked across rounds, sessions, modalities, or references. Y_3 Evolutionary Synthesis: evidence includes updates, conflicts, disappearance/reappearance, replacement, or overrides; the answer depends on the current valid state or on resolving a temporal change.

Boundary rules. If the main challenge is WHERE something is in a local area, prefer X_2. If the main challenge is WHICH specific instance is being referred to among similar objects or people, prefer X_3. If the answer requires reading exact text, exact color, small markings, or other fine attributes, prefer X_4. For Y,

use Y_3 only when later evidence changes or overrides earlier evidence; otherwise use Y_2 for non-conflicting multi-clue association and Y_1 for single-clue retrieval.

Output format. Return only a JSON object with fields `X`, `Y`, and `reasoning`. The labels must be one of `X1`, `X2`, `X3`, `X4` and one of `Y1`, `Y2`, `Y3`. Keep the reasoning to one or two sentences.

LLM-as-a-Judge Evaluation

You are an impartial judge evaluating the memory capabilities of an AI assistant on a question-answering task. Compare the Assistant's Answer against the Ground Truth and assign one score from $\{0, 0.25, 0.5, 0.75, 1\}$.

Important principles. Use semantic equivalence over surface form. Numeric and counting questions are binary: the exact number receives 1, and a wrong number receives 0. For negation-plus-correction questions, the answer must include both the correct polarity and the corrected fact. Identity questions are correct if the same unique person, object, or event is identified, even if the wording differs. Judge whether the answer gives the requested attribute or type of information. Wrong entities should not receive high scores just because they are on topic. Multi-item questions require matching the required set. Chronology and ordering questions require the correct ordered item. Do not over-reward fluent explanations. Opposite relations or orderings are normally scored 0.

Scoring rubric. Score 0 for contradictions, wrong yes/no polarity, wrong numbers, wrong answer type, wrong relation or order, hallucination, or no relevant information. Score 0.25 for answers that touch the topic but miss the core entity or value, contain major wrong associations, are excessively vague, or get most of a required set wrong. Score 0.5 for answers that capture the main entity or concept but miss important qualifiers, or for multi-item answers that get a substantial part of the set right. Score 0.75 for largely accurate answers with only minor missing detail or unnecessary filler. Score 1 for accurate, precise answers that contain all core information without hallucinations; exact wording is not required.

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

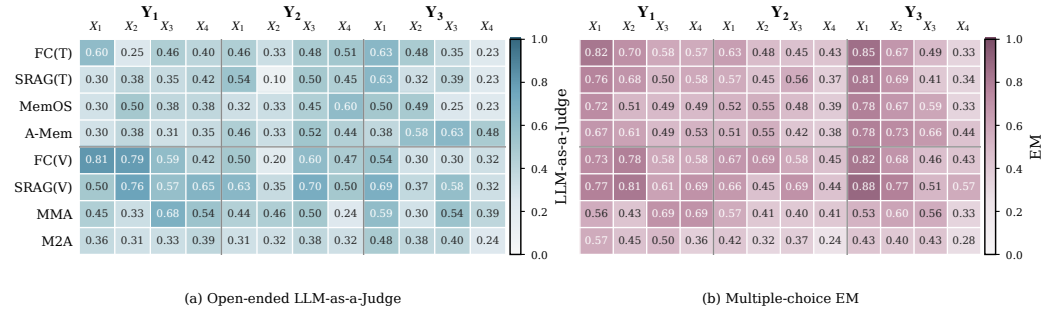
Output format. Return only a JSON object with fields `score` and `reasoning`. The score must be one of `0`, `0.25`, `0.5`, `0.75`, or `1`; the reasoning should be short.

C.4 Bootstrap Confidence Intervals

We use nonparametric bootstrap confidence intervals to characterize uncertainty in the aggregate comparisons that support the main claims. Unless otherwise stated, each interval is computed with 10,000 bootstrap resamples and a fixed random seed. For each resample, we sample original questions with replacement and recompute the target aggregate metric. For multiple-choice evaluation, each original question contributes its rotation-averaged EM score; the four answer rotations are treated as within-question variants rather than independent samples. For open-ended evaluation, each original question contributes its BLEU-1 or LLM-Judge score.

For paired comparisons, such as native-visual versus caption-only streams or two memory methods evaluated on the same question set, each bootstrap resample is drawn at the question level and the paired score difference is recomputed on the sampled questions. This preserves the correlation induced by evaluating different methods on the same benchmark items. We report percentile 95% confidence intervals. These intervals reflect sampling uncertainty over questions; they do not account for uncertainty from the judge model or judge prompt. We therefore use the intervals to assess the stability of aggregate trends, while treating individual small (X, Y) cells as diagnostic localization rather than independent significance claims.

Table 10 reports the key intervals used in the main analysis.



(a) Open-ended LLM-as-a-Judge (b) Multiple-choice EM

Figure 8: Representative method performance heatmap using `gpt-5.4-mini`. Left: LLM-as-a-Judge; Right: MCQ EM.

C.5 Memory Method Implementations

All methods are evaluated with the same benchmark runner. Non-agentic methods construct a method-specific context and pass it to the answer VLM. Agentic methods first ingest the full dialogue history into their memory store, then answer via the method-specific query interface. To maintain consistency, all internal VLM/LLM calls use the same backbone as the corresponding answering model, unless otherwise specified. Answer generation uses deterministic decoding with temperature 0 and the benchmark answer model for the run, unless otherwise specified.

We group methods into text-based and multimodal families. For text-based methods, images are replaced with dense captions generated by `gpt-5.2`. For multimodal methods, the original images are retained when supported by the method implementation. When official code is available and compatible with our benchmark runner, we use it with a lightweight adapter for MemEye evaluation formatting. Otherwise, we implement the method as described in the paper and report the relevant configuration below. All methods use the same benchmark prompts, decoding settings, and evaluation pipeline.

Full Context. FC(T) and FC(V) are full-history baselines. FC(T) concatenates the captioned dialogue history, while FC(V) passes the native multimodal history. When the estimated history exceeds the context budget, we apply FIFO (First-in-first-out) truncation. The context limit is 128k tokens, with reserved space for the question and answer.

Semantic RAG. SRAG(T) indexes each dialogue round with `all-MiniLM-L6-v2` text embeddings and retrieves the top-K=10 rounds. Image-bearing rounds are indexed using their captions. SRAG(V) uses the same text encoder and additionally embeds images with `siglip2-base-patch16-384`. Text and image similarities are combined with equal weights, and the selected top-K=10 rounds are passed to the answer VLM with native images.

A-Mem (Xu et al., 2026). A-Mem is evaluated as a text-only agentic memory system. Each dialogue round is converted into a memory note containing the dialog text and image caption lines. A-Mem uses `all-MiniLM-L6-v2` embeddings and retrieves K=10 related memories.

Reflexion (Shinn et al., 2023). Reflexion stores each captioned round as an observation in an RFMemory-style store. At query time, it recalls relevant memory text, answers using the shared benchmark prompt, and updates a compact global reflection. The recalled context is truncated to 6000 words.

Generative Agents (Park et al., 2023). Generative Agents stores captioned dialogue rounds as timestamped memories. Retrieval combines text similarity, recency, and importance using `all-MiniLM-L6-v2` embeddings, top-K=10 retrieval, a recency decay of 0.995, and a 4000-word recall budget. We enable the reflection mechanism with a threshold of 8.0, two reflection questions, and two generated insights.

MemoryOS (Kang et al., 2025). MemoryOS is evaluated through its official short-, mid-, and long-term memory interface. Each captioned round is added as a timestamped user/assistant memory. Retrieval uses the internal MemoryOS context retriever with short-term capacity 10, mid-term capacity 2000, long-term knowledge capacity 100, retrieval queue capacity 10, mid-term heat threshold 5.0, and mid-term similarity threshold 0.6. The embedding model is `all-MiniLM-L6-v2`.

SimpleMem (Liu et al., 2025; 2026). SimpleMem(T) uses the Omni-SimpleMem orchestrator in caption-based mode. Each round is inserted as a text memory with session, round, image-id, and timestamp metadata. We use `all-MiniLM-L6-v2` embeddings, retrieve K=20 memories, and disable SimpleMem's self-evolution loop for benchmark fairness. SimpleMem(V) uses the same orchestrator in multimodal mode. Retrieved memory units retain raw image pointers and load images at answer time.

MIRIX (Wang & Chen, 2025). MIRIX is evaluated through the official runtime adapter and agent wrapper. During memory construction, it receives dialogue messages with raw images and preserves message timestamps. MIRIX extracts information from the image inputs into structured memory entries rather than storing the raw images.

MMA (Lu et al., 2026). MMA is implemented as a confidence-aware multimodal memory system. It stores user and assistant turns as memory entries, attaches native images to user memories, computes text embeddings with `all-MiniLM-L6-v2`, and computes image embeddings with `google/siglip-so400m-patch14-384`. Retrieval ranks memories based on a blended text-image similarity score weighted by a confidence score. Confidence combines source credibility, temporal decay, and consensus with weights 0.45, 0.40, and 0.15, respectively, using a 30-day half-life. The retrieval budget is K=10.

M2A (Feng et al., 2026). M2A is evaluated with its two-phase protocol. During ingestion, sessions are processed in temporal order, and the MemoryManager writes semantic memories from raw dialogue and image evidence. During answering, a fresh query agent retrieves semantic memories and expands their evidence identifiers back to source rounds. We use `all-MiniLM-L6-v2` for text embeddings and `siglip2-base-patch16-384` for multimodal embeddings. The memory-manager loop allows up to 15 iterations with a recent-context window of 5 turns, and the query loop allows up to 5 iterations.

D Additional Results and Diagnostics

D.1 Additional Result Matrices

Tables 16, 17, 18, and 19 report coordinate-level results on a 371-question subset of MemEye for `gpt-5.4-mini`, `gpt-4.1-nano`, Qwen3-VL-8B-Instruct, and `gemini-2.5-flash-lite`, respectively. These tables include EM, BLEU-1, and LLM-as-a-Judge; the main `gpt-5.4-mini` results in Table 2 use the full 1,200-question benchmark.

D.2 Caption Robustness Ablation

The Caption-Proof gap reported in §4.4 uses generic dense captions generated by `gpt-5.2`. A natural concern is that the gap at high X may reflect captioner weakness rather than a structural modality boundary. To test this, we regenerate captions with a stronger captioner (`gpt-5.4-mini`) and a task-aware prompt that explicitly targets OCR content, exact colors, instance identities, spatial relations, and fine-grained visual attributes. The prompt instructs the model to be exhaustive about details that a generic caption might omit, producing captions that are typically 2–3$\times$ longer than the originals.

We evaluate on a stratified subsample of 120 open-ended questions (40 low-X from X_1–X_2, 80 high-X from X_3-X_4), drawn by round-robin across all eight tasks. We run FC(T) and SRAG(T) with the task-aware captions and compare against both the original generic-caption runs and the visual-stream runs on the same questions.

Table 11: Caption robustness ablation on `gpt-5.4-mini` (LLM-Judge, open-ended). Rows denote matched memory families. Native-image input reports the multimodal version, FC(V) or SRAG(V), while caption input reports the corresponding caption-based version, FC(T) or SRAG(T), using either generic or task-aware captions. Δ is the gain from native images over each caption condition. The stratified subsample contains 40 low-X and 80 high-X questions across all eight tasks.

Method family	X group	Native-image input	Caption input		Δ (Native image − Caption)	
			Generic	Task-aware	Generic	Task-aware
FC	Low-X	0.449	0.528	0.636	−0.079	−0.188
FC	High-X	0.432	0.298	0.418	+0.134	+0.014
FC	X_4	0.485	0.308	0.386	+0.177	+0.100
SRAG	Low-X	0.590	0.425	0.595	+0.165	−0.005
SRAG	High-X	0.428	0.235	0.387	+0.194	+0.041
SRAG	X_4	0.419	0.203	0.325	+0.215	+0.094

Table 11 reports the results. Task-aware captions substantially improve textual-stream performance: SRAG(T) rises from 0.425 to 0.595 at low X and from 0.235 to 0.387 at high X. Critically, the visual-stream gain shows a diminishing-returns pattern. At low X, the gap closes entirely: SRAG(T) with task-aware captions matches or exceeds the visual stream ($\Delta = -0.005$). At high X, the gap shrinks from $+0.194$ to $+0.041$ but remains positive; at X_4 it shrinks from $+0.215$ to $+0.094$. This pattern suggests that scene- and region-level evidence can often be recovered by a sufficiently detailed captioner, while instance- and pixel-level evidence—identity bindings, fine textures, small text, exact color attributes—remains harder to fully recover through textual mediation, even with a stronger task-aware captioning prompt.

D.3 Effective Visual Information Analysis

Figure 2 compares how much useful visual information remains under different input settings across LoCoMo (Maharana et al., 2024), MMRC (Xue et al., 2025), Mem-Gallery (Bei et al., 2026), and MemEye. The goal of this analysis is not to evaluate long-context retrieval, but to test whether the answer-relevant image information in each benchmark can be replaced by text. To isolate this factor, we use the gold clue rounds for each question whenever clue annotations are available. This removes memory search as the primary bottleneck and focuses the comparison on how much information native images contribute beyond textual context or captions.

For each benchmark, we evaluate the same answering model, `gpt-5.4-mini`, under three input settings. In the *No Visual Info.* setting, the model receives the question and the gold textual clue context, but no images or image captions. In the *Caption Only* setting, each image in the gold clue rounds is replaced by a dense textual caption. In the *Multimodal* setting, the model receives the same gold clue rounds with the original images. The question text is kept fixed across all three settings, so differences in performance reflect the additional information provided by captions or native visual inputs.

For MemEye, we use the annotated clue rounds provided by the benchmark. For prior benchmarks, we construct comparable gold-clue contexts from their provided answer-relevant turns, evidence annotations, or, when available, a minimal supporting dialogue context. Captions are generated using `gpt-5.2` with the same dense captioning prompt used in our caption-based memory experiments. All answers are evaluated in the open-ended format using the same LLM-as-a-Judge protocol as the main experiments, and scores are multiplied by 100 for visualization.

This analysis measures visual irreplaceability. If caption-only performance approaches multimodal performance, then the benchmark's visually grounded questions are largely recoverable from textual substitutions. If multimodal performance substantially exceeds caption-only performance, then native image evidence contains information that captions do not preserve. As shown in Figure 2, MemEye exhibits a larger gain from caption-only to multimodal input than the prior benchmarks, indicating that its questions contain more irreplaceable visual evidence.

Table 12: Clue-round retrieval diagnostics on open-ended `gpt-5.4-mini` runs. Metrics compare the retrieved top-10 memory rounds with annotated gold clue rounds. The sample count n reflects questions with valid retrieved-round provenance for this diagnostic and can differ slightly from axis-level dataset totals.

Method	Y	n	Any-Clue	Coverage	Full-Clue	Latest / Stale
SRAG(T)	Y_1	113	0.832	0.819	0.805	– / –
SRAG(T)	Y_2	195	0.826	0.590	0.344	– / –
SRAG(T)	Y_3	60	0.667	0.510	0.367	0.517 / 0.526
SRAG(V)	Y_1	113	0.885	0.870	0.858	– / –
SRAG(V)	Y_2	195	0.826	0.622	0.410	– / –
SRAG(V)	Y_3	60	0.750	0.553	0.367	0.533 / 0.767
MMA	Y_1	113	0.841	0.824	0.805	– / –
MMA	Y_2	195	0.831	0.612	0.395	– / –
MMA	Y_3	60	0.833	0.583	0.383	0.550 / 0.750
M2A	Y_1	113	0.319	0.295	0.274	– / –
M2A	Y_2	195	0.477	0.234	0.051	– / –
M2A	Y_3	60	0.267	0.124	0.067	0.133 / 0.500

D.4 Retrieval Diagnostics

We use the annotated clue rounds to diagnose retrieval failures after a question is asked. The goal is not to build a separate retrieval leaderboard, but to separate three sources of error: a system may fail to retrieve relevant evidence, retrieve only part of the required evidence chain, or retrieve evidence that is semantically relevant but temporally stale. For retrieval-based methods, we compare the top-K retrieved rounds with the gold clue rounds. *Any-Clue Recall@K* measures whether at least one gold clue is retrieved. *Coverage@K* measures the fraction of gold clues retrieved. *Full-Clue Recall@K* measures whether the complete clue set is retrieved. For Y_3 questions, we additionally report *Latest-Clue Recall@K*, which checks whether the final decisive clue is retrieved, and *Stale-Dominance*, which measures whether stale evidence is ranked above the latest clue or appears without the latest clue.

Table 12 shows that native visual retrieval improves evidence access, but does not solve evolving-state reasoning. SRAG(V) improves Any-Clue Recall@10 over SRAG(T), especially at Y_1 and Y_3. However, on Y_3, its Full-Clue Recall@10 is only 0.367 and its Latest-Clue Recall@10 is only 0.533. Thus, many failures are not caused by a complete absence of relevant evidence. Instead, the system often retrieves an incomplete update chain or misses the clue that establishes the current state. MMA retrieves Y_3 clue evidence slightly more often than SRAG(V), but under the broader source-session stale definition used by the recency probe, its Stale-Dominance remains high (0.750 versus 0.767 for SRAG(V)). M2A should be interpreted carefully because its provenance is expanded from semantic-memory evidence rather than direct raw-round retrieval. Overall, these diagnostics identify a specific failure mode: semantic relevance can surface the right topic while still selecting an outdated visual state.

We next test whether stale-state selection can be reduced by a simple recency signal. Table 13 keeps SRAG(V)'s retrieved candidate pool fixed and re-ranks candidates using a mixture of retrieval similarity and exponential recency:

$$s_i = \alpha \, \mathrm{sim}_i + (1 - \alpha) \exp(-\lambda \Delta t_i),$$

where sim_i is the SigLIP2 retrieval similarity and Δt_i is the age of candidate round i relative to the query. We set $\lambda = 0.02$. The age Δt_i is measured as dialogue-round distance between candidate round i and the query round, using the logged `recency_age` field, with no additional normalization. This is a retrieval-side diagnostic probe rather than a new memory method: the candidate pool is unchanged, so recency can correct stale-over-latest ranking errors but cannot recover a latest clue that was not retrieved.

We decompose stale evidence selection into two cases. *Rank-Inversion* measures how often both stale and latest evidence are retrieved but stale evidence is ranked higher. *Latest-Miss* measures how often the latest decisive evidence is absent from the retrieved pool. The recency probe reduces

Table 13: Recency counterfactual for SRAG(V) on Y_3 open-ended questions. Stale-Dominance uses the source-session stale definition: any retrieved round from a source session that occurs before the latest annotated clue is treated as stale evidence. Latest-Miss measures cases where the latest clue is absent from the top-10 candidate set; Rank-Inversion measures cases where the latest clue is retrieved but stale evidence ranks above it. Paired deltas are relative to vanilla SRAG(V), with percentile 95% bootstrap CIs over 10,000 question-level resamples. Y3 Judge is reported only for settings with full answer regeneration.

Ranking	Latest@10	Latest-Miss	Stale-Dom.	Rank-Inv.	Y3 Judge
SRAG(V), α=1.0	0.600	0.400	0.767	0.483	0.292
+ Recency, α=0.7	0.550	0.450	0.683	0.367	0.358
+ Recency, α=0.5	0.550	0.450	0.583	0.283	0.325
Paired deltas relative to SRAG(V)					
Ranking	Δ Stale-Dom.	95% CI	Δ Rank-Inv.	95% CI	Δ Judge
+ Recency, α=0.7	-0.083	$[-0.167, -0.017]$	-0.117	$[-0.217, -0.033]$	$+0.067$
+ Recency, α=0.5	-0.183	$[-0.283, -0.100]$	-0.200	$[-0.317, -0.083]$	$+0.033$

Table 14: Answer-regeneration sanity check for the SRAG(V) recency diagnostic. Values are open-ended LLM-as-a-Judge scores. Δ reports paired differences relative to SRAG(V), with percentile 95% bootstrap CIs over 10,000 question-level resamples.

Setting	Slice	Judge	Δ	95% CI
SRAG(V)	Y_1	0.673	–	–
SRAG(V)	Y_2	0.449	–	–
SRAG(V)	Y_3	0.292	–	–
SRAG(V)	Avg.	0.492	–	–
+ Recency α=0.7	Y_1	0.677	$+0.004$	$[-0.058, +0.066]$
+ Recency α=0.7	Y_2	0.399	-0.051	$[-0.106, +0.005]$
+ Recency α=0.7	Y_3	0.358	$+0.067$	$[-0.042, +0.175]$
+ Recency α=0.7	Avg.	0.477	-0.015	$[-0.054, +0.024]$
+ Recency α=0.5	Y_1	0.650	-0.022	$[-0.093, +0.049]$
+ Recency α=0.5	Y_2	0.391	-0.058	$[-0.129, +0.013]$
+ Recency α=0.5	Y_3	0.325	$+0.033$	$[-0.075, +0.142]$
+ Recency α=0.5	Avg.	0.460	-0.032	$[-0.080, +0.015]$

Stale-Dominance and Rank-Inversion, showing that part of the Y_3 failure comes from ranking stale evidence above the valid update. However, Latest-Miss does not improve, because re-ranking cannot recover missing evidence. Table 14 reports answer-regeneration checks for α=0.7 and α=0.5. The Y_3 LLM-Judge point estimates improve, but the confidence intervals overlap zero, so the aggregate answer-quality gains are not reliable.

The two recency weights show the same trade-off. The α=0.7 setting yields the larger Y_3 answer-quality point estimate ($+0.067$ Judge), while α=0.5 yields the larger retrieval-side reduction in Stale-Dominance (-0.183 versus -0.083). Thus, recency is useful as a diagnostic signal for temporal-authority failures, but it is not a complete solution. A memory system must also retrieve the missing latest evidence and reason over the update chain once the evidence is available.

D.5 Evolving Visual State Probe

The retrieval diagnostics above show that Y_3 failures involve stale evidence selection. We now isolate the visual form of this problem. In the evolving visual-state subset, the update is expressed in the images themselves rather than only in dialogue text. The questions require a sequence of visual clues in which later images update, conflict with, or override earlier images. To answer correctly, a memory system must recover the relevant visual evidence and determine which visual state remains valid.

Table 15: Evolving visual-state probe on the subset of evolving visual state Y_3 questions under `gpt-5.4-mini`. Scores are open-ended LLM-as-a-Judge averages. Oracle settings use clue-only evidence with raw images. Memory-system settings use each method's original retrieval or context behavior.

Evaluation	Method	LLM-as-a-Judge
Oracle control	Stale-only	0.591
	Latest-only	0.712
	All-clue oracle	0.727
Memory system	FC(V)	0.333
	SRAG(V)	0.379
	MMA	0.394
	M2A	0.182

We evaluate three controlled evidence settings. *Latest-only* provides only the latest decisive visual clue and tests whether the current state is visually readable. *Stale-only* provides only earlier visual clues and tests whether outdated evidence supports a plausible but stale answer. *All-clue oracle* provides the full annotated visual evidence chain and tests whether the answering model can use both stale and updated clues when evidence selection is perfect. We compare these controls with representative multimodal memory systems under their normal retrieval or context behavior.

Table 15 reports the LLM-as-a-Judge scores. The all-clue oracle obtains the highest score, showing that the questions become substantially more answerable when the full visual update chain is provided. Latest-only is slightly lower, as expected, because some questions ask about the change itself and therefore require earlier visual states for comparison. Stale-only is lower than both latest-only and all-clue oracle, indicating that older visual states are not harmless context; they can support outdated or incomplete answers. All memory systems remain far below the controlled evidence settings. This gap shows that current multimodal memory systems struggle not only to store images, but also to retrieve and prioritize the valid visual state when stale and updated visual evidence coexist.

This failure mode is difficult to expose in benchmarks where updates are primarily textual. In MemEye, the key challenge is temporal authority over native visual evidence: the system must decide which image-grounded state is current. Figure 12 provides three case studies illustrating this behavior. The cases show that retrieval failures are not uniform: a method may miss the latest clue, over-weight stale evidence, or retrieve locally plausible fragments without tracking the full temporal chain.

D.6 Cross-topic Dialogue Scaling Ablation

Figure 5(d) and §4.5 summarize the cross-topic dialogue scaling ablation in the main text. Here we describe the construction protocol and full trends. The goal is to test whether memory systems remain robust when answer-relevant evidence is embedded in a larger, more diverse history containing unrelated tasks.

We construct two controlled four-domain combinations. The first combines Brand Memory, Social Chat, Cartoon Entertainment, and Card Playlog; the second combines Health Care, Home Renovation, CrossScene Memory, and Outdoor Navigation. For each combination, we evaluate three memory scales over the same underlying questions. The 1-dataset point is the QA-weighted average over the four clean single-domain runs. The 2-dataset point is the QA-weighted average over two pairwise combinations. The 4-dataset point evaluates the full combined conversation history. This design keeps the answer-relevant evidence fixed while increasing cross-domain interference, allowing us to isolate whether a memory system can route the correct multimodal evidence from a noisy long-horizon history.

Figure 9 shows the full scaling curves for both MCQ EM and open-ended LLM-as-a-Judge. The detailed trends support the main-text interpretation. Full Context (V) is sensitive to cross-topic interference: its performance drops as unrelated histories are added, especially in the Health Care, Home Renovation, CrossScene Memory, and Outdoor Navigation combination. This decline likely reflects two factors. First, the model must filter more irrelevant visual-textual evidence as the history

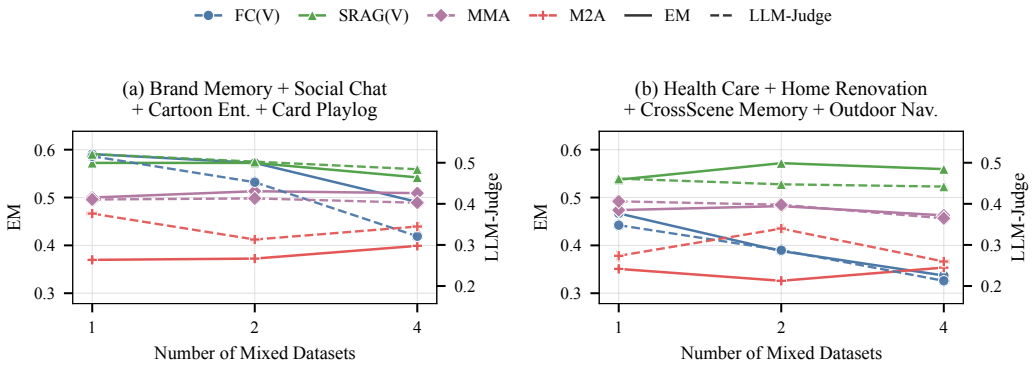


Figure 9: Cross-topic dialogue scaling ablation under `gpt-5.4-mini`. Left: MCQ EM. Right: LLM-as-a-Judge. Each curve evaluates the same questions as memory scale increases from clean single-domain histories, to pairwise cross-domain histories, to the full four-domain history. Results are averaged over two controlled four-domain combinations.

grows. Second, long histories may approach the context window, increasing the risk that early answer-relevant evidence is truncated or diluted.

Retrieval-based and structured memory methods are more stable. Semantic RAG (V) remains comparatively robust across memory scales, suggesting that targeted retrieval reduces irrelevant context before answer generation. MMA also exhibits flatter trends, indicating that structured multimodal memory can reduce interference. M2A shows partial resilience in open-ended evaluation, suggesting that agentic operations such as memory writing, conflict checking, or iterative querying can help manage noisy histories. However, its lower absolute scores indicate that abstraction must still preserve fine-grained visual evidence to be effective.

Overall, this ablation shows that scaling multimodal memory is not simply a matter of increasing the context window. Long-term multimodal agents need evidence-routing mechanisms that filter unrelated history while preserving the visual details and state information needed for the current question.

D.7 Limitations and Broader Impacts

MemEye is intended as a diagnostic benchmark rather than a complete simulation of all multimodal agent deployments. Its scenarios, model panel, captioning pipeline, and human validation sample may not cover every real-world setting. Comparisons across memory architectures can also be affected by method-specific encoders and implementations, so we interpret the results as diagnostic evidence about design trade-offs rather than as a universal ranking of memory systems. Finally, the human oracle check in Appendix B.3 is a small sanity check rather than a full human ceiling estimate.

Evaluating long-term visual memory also has broader implications. More reliable memory can reduce stale or incorrect visual-state reasoning, but stronger visual memory systems may increase privacy risks when agents store user images, personal environments, or evolving user states. Future multimodal memory systems should therefore combine evidence-preserving memory with consent, data minimization, deletion, and access-control mechanisms.

Table 16: Results on a subset of the MemEye evaluation matrix using `gpt-5.4-mini`. Columns correspond to memory methods, grouped into text-only and multimodal families. Within each coordinate, EM is reported for multiple-choice questions, while BLEU-1 and LLM-as-a-Judge (LLM-Judge) is reported for free-response questions. The first- and second-performing memory model(s) are highlighted with orange and blue backgrounds, respectively.

Y	X	Metric	Textual memory							Multimodal memory					
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MIRIX	MMA	M2A	SM(V)
Y1	X1	EM	0.8000	0.9500	0.6750	0.2500	0.7750	0.7750	0.8000	1.0000	0.9000	0.6750	0.5500	0.4750	0.8500
		BLEU-1	0.3291	0.2654	0.2270	0.1914	0.1596	0.1312	0.2926	0.3867	0.3814	0.1178	0.2539	0.1439	0.2922
		LLM-Judge	0.6500	0.4500	0.6000	0.3000	0.4000	0.2500	0.5500	0.6500	0.6000	0.4000	0.4500	0.4500	0.5000
	X2	EM	0.7500	0.5455	0.2045	0.2273	0.4545	0.4318	0.5000	0.6818	0.7727	0.5227	0.5455	0.2500	0.4773
		BLEU-1	0.2282	0.2819	0.0894	0.0829	0.3242	0.3333	0.1220	0.2909	0.6076	0.2144	0.3471	0.1738	0.0766
		LLM-Judge	0.4545	0.4091	0.1364	0.1818	0.4091	0.3182	0.2727	0.5455	0.9091	0.5000	0.7273	0.3182	0.1818
	X3	EM	0.4662	0.4527	0.2534	0.2500	0.3649	0.4392	0.3209	0.5709	0.6554	0.5473	0.5507	0.3615	0.3784
		BLEU-1	0.2692	0.2675	0.2196	0.1685	0.2547	0.3142	0.1109	0.3597	0.3918	0.1810	0.3554	0.1889	0.1532
		LLM-Judge	0.3716	0.3176	0.3108	0.2230	0.3176	0.4459	0.2230	0.5338	0.6554	0.2568	0.5946	0.3311	0.2838
	X4	EM	0.4722	0.5694	0.4444	0.2361	0.6389	0.5972	0.4583	0.4722	0.8056	0.3750	0.6250	0.5000	0.5000
		BLEU-1	0.2861	0.2907	0.0722	0.0327	0.2240	0.1852	0.1139	0.1676	0.3423	0.1321	0.3237	0.0658	0.1139
		LLM-Judge	0.3889	0.3333	0.1111	0.0556	0.2222	0.2500	0.2222	0.3333	0.6389	0.3056	0.6389	0.1667	0.2222
Y2	X1	EM	0.5086	0.5172	0.4483	0.2500	0.4569	0.4052	0.5000	0.5345	0.5086	0.2845	0.5086	0.3190	0.5690
		BLEU-1	0.3851	0.5148	0.4369	0.2195	0.3098	0.3104	0.4071	0.4787	0.5661	0.3748	0.3101	0.0848	0.3381
		LLM-Judge	0.4483	0.5345	0.5517	0.2759	0.3793	0.4138	0.4828	0.5000	0.6552	0.4138	0.3621	0.2931	0.4138
	X2	EM	0.4881	0.3810	0.1905	0.2619	0.3333	0.3690	0.3095	0.3810	0.2976	0.3214	0.4762	0.3333	0.3452
		BLEU-1	0.1208	0.1045	0.1404	0.0469	0.0689	0.1045	0.1254	0.0996	0.1590	0.0749	0.1562	0.0989	0.0702
		LLM-Judge	0.1667	0.0952	0.1667	0.0238	0.1429	0.0952	0.1429	0.1667	0.2143	0.1429	0.1905	0.2143	0.0952
	X3	EM	0.4417	0.4125	0.3833	0.2292	0.3333	0.3917	0.3917	0.5750	0.6250	0.4583	0.4292	0.3792	0.3708
		BLEU-1	0.2018	0.1800	0.1255	0.1368	0.1582	0.1759	0.0879	0.1694	0.2532	0.1259	0.1863	0.1174	0.1322
		LLM-Judge	0.3750	0.2833	0.2000	0.1167	0.2833	0.2917	0.2333	0.3917	0.6000	0.2667	0.3750	0.3583	0.2667
	X4	EM	0.3438	0.3523	0.2841	0.2472	0.3409	0.3551	0.3097	0.3636	0.3722	0.3665	0.4119	0.2955	0.2869
		BLEU-1	0.3487	0.2786	0.3381	0.2161	0.3904	0.2305	0.2321	0.3496	0.2737	0.2827	0.2425	0.1082	0.1923
		LLM-Judge	0.3807	0.3182	0.3466	0.2614	0.4205	0.3636	0.2500	0.3977	0.3352	0.3352	0.2898	0.3352	0.2159
Y3	X1	EM	0.8000	0.7500	0.7000	0.2500	0.7000	0.4750	0.5000	0.9000	0.5750	0.6750	0.5500	0.6000	0.5500
		BLEU-1	0.2037	0.2190	0.3152	0.1832	0.2106	0.2107	0.2212	0.3097	0.1251	0.2675	0.2791	0.1311	0.2323
		LLM-Judge	0.6500	0.6000	0.7000	0.4500	0.6000	0.5000	0.6000	0.6500	0.4500	0.6000	0.5500	0.6000	0.6000
	X2	EM	0.7000	0.7000	0.5500	0.2500	0.4750	0.6000	0.6500	0.6000	0.7750	0.4750	0.6750	0.4750	0.7000
		BLEU-1	0.2582	0.2936	0.2961	0.2592	0.2065	0.3395	0.2492	0.2097	0.1978	0.1045	0.1819	0.1126	0.2502
		LLM-Judge	0.4000	0.5000	0.4500	0.4500	0.3000	0.4500	0.3000	0.2500	0.3500	0.2000	0.3000	0.1000	0.3000
	X3	EM	0.6000	0.6250	0.5750	0.2750	0.6250	0.5750	0.5250	0.5750	0.6500	0.6500	0.8000	0.5000	0.6000
		BLEU-1	0.1259	0.1468	0.1152	0.2134	0.1995	0.2275	0.0523	0.1380	0.1078	0.1365	0.1738	0.1783	0.1485
		LLM-Judge	0.5500	0.5500	0.5500	0.4000	0.3500	0.5500	0.3500	0.4000	0.3000	0.4000	0.4500	0.5500	0.3500
	X4	EM	0.4333	0.3250	0.2833	0.2333	0.3417	0.3417	0.3167	0.5917	0.4750	0.2583	0.3417	0.3250	0.3083
		BLEU-1	0.2122	0.2630	0.2463	0.2407	0.1096	0.1302	0.1969	0.3495	0.1613	0.1477	0.2708	0.0690	0.2136
		LLM-Judge	0.3000	0.3000	0.2833	0.3167	0.1667	0.3000	0.2000	0.4500	0.2167	0.1667	0.2667	0.3000	0.2000
Avg.	–	EM	0.5670	0.5484	0.4160	0.2467	0.4866	0.4797	0.4651	0.6038	0.6177	0.4674	0.5386	0.4011	0.4947
		BLEU-1	0.2474	0.2588	0.2185	0.1659	0.2180	0.2244	0.1843	0.2758	0.2972	0.1800	0.2567	0.1227	0.1844
		LLM-Judge	0.4280	0.3909	0.3672	0.2546	0.3326	0.3524	0.3189	0.4391	0.4937	0.3323	0.4329	0.3347	0.3025

Table 17: Results on a subset of the MemEye evaluation matrix using `gpt-4.1-nano`. Columns correspond to memory methods, grouped into text-only and multimodal families. Within each coordinate, EM is reported for multiple-choice questions, while BLEU-1 and LLM-as-a-Judge (LLM-Judge) are reported for free-response questions. The first- and second-performing memory model(s) are highlighted with orange and blue backgrounds, respectively.

Y	X	Metric	Textual memory							Multimodal memory					
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MIRIX	MMA	M2A	SM(V)
Y1	X1	EM	0.4750	0.7500	0.7500	0.2500	0.7250	0.7000	0.6750	0.6750	0.7750	0.4750	0.6000	0.3500	0.7500
		BLEU-1	0.2377	0.1875	0.2991	0.1168	0.1358	0.1540	0.2008	0.1523	0.2506	0.0594	0.1333	0.0467	0.2579
		LLM-Judge	0.4500	0.3500	0.5500	0.2500	0.4000	0.4500	0.5500	0.4500	0.6500	0.2000	0.3000	0.1000	0.5000
	X2	EM	0.4444	0.4722	0.4167	0.2222	0.4167	0.5556	0.5833	0.3333	0.2778	0.3056	0.2500	0.5556	0.6111
		BLEU-1	0.2100	0.1838	0.2874	0.1421	0.1295	0.1011	0.0428	0.0921	0.2364	0.0431	0.0960	0.0657	0.0314
		LLM-Judge	0.3333	0.3333	0.3333	0.3333	0.3333	0.1111	0.3333	0.3889	0.6111	0.3889	0.3889	0.1111	0.3333
	X3	EM	0.3125	0.2961	0.3388	0.2599	0.2961	0.2632	0.2796	0.3947	0.4342	0.4079	0.4079	0.2796	0.2697
		BLEU-1	0.2341	0.2007	0.2106	0.1328	0.2069	0.1869	0.1644	0.2275	0.2480	0.0781	0.1598	0.0724	0.2023
		LLM-Judge	0.2763	0.2303	0.2895	0.2237	0.2632	0.2039	0.2039	0.3882	0.5263	0.1645	0.4868	0.1118	0.2237
	X4	EM	0.4167	0.5139	0.4167	0.2778	0.5694	0.5556	0.3611	0.4861	0.5972	0.5556	0.5556	0.3472	0.3611
		BLEU-1	0.1283	0.2408	0.1022	0.0148	0.2018	0.1963	0.1013	0.1032	0.1259	0.0197	0.0608	0.0356	0.0694
		LLM-Judge	0.2778	0.4444	0.1667	0.0000	0.2778	0.3056	0.1111	0.2222	0.3611	0.0000	0.3611	0.1389	0.1667
Y2	X1	EM	0.3534	0.3966	0.3534	0.2414	0.3017	0.3190	0.2931	0.4052	0.4052	0.2069	0.3879	0.3276	0.3448
		BLEU-1	0.1246	0.2092	0.2119	0.0503	0.2144	0.1232	0.1600	0.0815	0.1181	0.0151	0.0778	0.0229	0.2617
		LLM-Judge	0.3621	0.5000	0.4310	0.2586	0.4138	0.3621	0.2241	0.5000	0.3621	0.1034	0.3621	0.1034	0.3276
	X2	EM	0.3910	0.2949	0.3718	0.2500	0.2692	0.3077	0.4295	0.3782	0.4551	0.3910	0.3782	0.3462	0.4744
		BLEU-1	0.1521	0.1421	0.1426	0.1129	0.1396	0.0947	0.1192	0.1400	0.1599	0.0785	0.1261	0.0988	0.0875
		LLM-Judge	0.3205	0.2436	0.1923	0.2436	0.1282	0.1667	0.2179	0.2308	0.4103	0.1795	0.2308	0.1538	0.2051
	X3	EM	0.3512	0.3393	0.3155	0.2381	0.3571	0.2917	0.3095	0.3631	0.4048	0.3988	0.3512	0.2976	0.2976
		BLEU-1	0.1368	0.0871	0.1347	0.0993	0.1715	0.1715	0.0843	0.1448	0.1676	0.0420	0.0945	0.0534	0.1322
		LLM-Judge	0.2500	0.1905	0.1429	0.1429	0.2262	0.2381	0.1667	0.3571	0.4048	0.0833	0.2143	0.0833	0.1667
	X4	EM	0.2784	0.2756	0.2642	0.2642	0.2699	0.2386	0.2756	0.2585	0.3324	0.2699	0.3438	0.3125	0.3182
		BLEU-1	0.2114	0.1705	0.1580	0.0760	0.1988	0.1542	0.2391	0.1742	0.1233	0.0345	0.0670	0.0566	0.2955
		LLM-Judge	0.3295	0.3125	0.2784	0.3580	0.3580	0.3295	0.2330	0.2386	0.3409	0.0966	0.2955	0.2727	0.3182
Y3	X1	EM	0.5750	0.5750	0.5000	0.2500	0.6500	0.4500	0.4250	0.5750	0.4500	0.5000	0.3500	0.3250	0.4750
		BLEU-1	0.2579	0.1544	0.2705	0.1493	0.1269	0.3162	0.1423	0.1519	0.1052	0.0295	0.1243	0.0487	0.1541
		LLM-Judge	0.5500	0.4000	0.6000	0.4000	0.2500	0.4500	0.4500	0.4500	0.3000	0.1000	0.4500	0.1000	0.3500
	X2	EM	0.6000	0.7000	0.5000	0.2500	0.4500	0.4500	0.6250	0.4500	0.4000	0.5000	0.2500	0.3500	0.6000
		BLEU-1	0.3175	0.3053	0.2436	0.2180	0.1376	0.3470	0.2635	0.2366	0.1936	0.1935	0.1309	0.0866	0.2903
		LLM-Judge	0.8000	0.7000	0.6000	0.7500	0.2500	0.7000	0.6500	0.5000	0.5000	0.3000	0.3000	0.2000	0.6500
	X3	EM	0.4250	0.5500	0.4500	0.3000	0.5500	0.4750	0.5500	0.4750	0.4250	0.7000	0.6000	0.3750	0.5750
		BLEU-1	0.1182	0.1960	0.1965	0.1255	0.1773	0.1177	0.1497	0.1105	0.2074	0.0580	0.1585	0.0809	0.1409
		LLM-Judge	0.3000	0.2500	0.2500	0.2000	0.2000	0.2500	0.1500	0.2500	0.4000	0.0000	0.2500	0.2000	0.2000
	X4	EM	0.1250	0.1833	0.1083	0.2417	0.2333	0.0833	0.1917	0.1417	0.1667	0.2000	0.2000	0.2250	0.1667
		BLEU-1	0.0579	0.1514	0.0607	0.0429	0.0612	0.0845	0.0888	0.1125	0.0528	0.0281	0.0416	0.0294	0.1009
		LLM-Judge	0.3333	0.3167	0.2333	0.1667	0.2167	0.2667	0.1000	0.2333	0.2000	0.0167	0.2000	0.0833	0.1000
Avg.	–	EM	0.3956	0.4456	0.3988	0.2538	0.4240	0.3908	0.4165	0.4113	0.4269	0.4092	0.3895	0.3409	0.4370
		BLEU-1	0.1822	0.1857	0.1931	0.1067	0.1585	0.1706	0.1463	0.1439	0.1657	0.0566	0.1059	0.0582	0.1687
		LLM-Judge	0.3819	0.3559	0.3390	0.2772	0.2764	0.3195	0.2825	0.3508	0.4222	0.1361	0.3200	0.1382	0.2951

Table 18: Results on a subset of the MemEye evaluation matrix using Qwen3-VL-8B-Instruct. Columns correspond to memory methods, grouped into text-only and multimodal families. Within each coordinate, EM is reported for multiple-choice questions, while BLEU-1 and LLM-as-a-Judge (LLM-Judge) are reported for free-response questions. The first- and second-performing memory model(s) are highlighted with orange and blue backgrounds, respectively.

Y	X	Metric	Textual memory							Multimodal memory				
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MMA	M2A	SM(V)
Y1	X1	EM	0.8500	0.8500	0.9250	0.2500	0.7750	0.8000	0.5250	0.5750	0.5000	0.5500	0.5250	0.5500
		BLEU-1	0.2658	0.1958	0.1976	0.1131	0.2471	0.1537	0.2425	0.1177	0.1093	0.1022	0.0874	0.2925
		LLM-Judge	0.4500	0.5500	0.4500	0.2500	0.5500	0.4000	0.5000	0.3000	0.3500	0.1500	0.1500	0.4500
	X2	EM	0.7500	0.5278	0.6389	0.2500	0.5278	0.5833	0.6111	0.3056	0.4444	0.2222	0.5278	0.6111
		BLEU-1	0.1561	0.1123	0.1471	0.0736	0.0874	0.2119	0.0150	0.1544	0.1123	0.0069	0.0712	0.0150
		LLM-Judge	0.6111	0.3333	0.3333	0.2778	0.1111	0.6111	0.2222	0.2222	0.2778	0.0000	0.1111	0.2222
	X3	EM	0.4901	0.4901	0.4605	0.2599	0.4243	0.4507	0.1118	0.3026	0.3586	0.3783	0.3717	0.1118
		BLEU-1	0.2823	0.2570	0.2737	0.1893	0.1690	0.1753	0.0589	0.1792	0.2719	0.2050	0.1192	0.0688
		LLM-Judge	0.3092	0.3158	0.3487	0.2303	0.2171	0.2632	0.1053	0.2500	0.2632	0.2237	0.2039	0.1118
	X4	EM	0.5278	0.6111	0.5278	0.2500	0.5556	0.5833	0.2639	0.4167	0.4306	0.2083	0.2917	0.2639
		BLEU-1	0.2051	0.2097	0.1046	0.0273	0.1525	0.1721	0.0556	0.0779	0.1771	0.1378	0.0377	0.0556
		LLM-Judge	0.3889	0.3611	0.2778	0.0833	0.2222	0.2778	0.1111	0.3333	0.3056	0.2222	0.0833	0.1111
Y2	X1	EM	0.4138	0.4655	0.4310	0.2414	0.3966	0.4052	0.3621	0.2241	0.4397	0.4741	0.3103	0.3707
		BLEU-1	0.2811	0.3167	0.3317	0.1716	0.1968	0.2836	0.2634	0.1179	0.4219	0.2108	0.0737	0.1939
		LLM-Judge	0.3621	0.3621	0.4310	0.2414	0.2414	0.3793	0.3276	0.2069	0.4828	0.2931	0.1034	0.2414
	X2	EM	0.4679	0.3397	0.4359	0.2564	0.3654	0.4038	0.3526	0.1923	0.4103	0.4167	0.4295	0.3397
		BLEU-1	0.2243	0.1091	0.2341	0.2144	0.0576	0.0785	0.0807	0.1542	0.1995	0.1163	0.0909	0.0765
		LLM-Judge	0.3590	0.2179	0.3205	0.2692	0.1538	0.1410	0.1538	0.1026	0.1795	0.1154	0.0513	0.1410
	X3	EM	0.3690	0.3452	0.3095	0.2500	0.3393	0.3155	0.1964	0.2024	0.2976	0.2679	0.3750	0.2262
		BLEU-1	0.1373	0.1174	0.0967	0.1155	0.1423	0.1303	0.0432	0.1012	0.1376	0.0303	0.0933	0.0457
		LLM-Judge	0.2500	0.2143	0.1548	0.1310	0.2381	0.1786	0.0952	0.1310	0.2024	0.0000	0.2143	0.1310
	X4	EM	0.3807	0.3693	0.3494	0.2500	0.3892	0.4034	0.1506	0.2756	0.3352	0.2869	0.3040	0.1307
		BLEU-1	0.2880	0.2265	0.2654	0.2712	0.2153	0.1769	0.0795	0.2097	0.1744	0.1898	0.0476	0.1003
		LLM-Judge	0.3239	0.2443	0.3011	0.2955	0.2216	0.2102	0.1023	0.2557	0.2102	0.2159	0.1250	0.1364
Y3	X1	EM	0.5750	0.5250	0.5250	0.2500	0.7500	0.7250	0.5250	0.2750	0.2250	0.2750	0.4250	0.4750
		BLEU-1	0.1623	0.1427	0.1667	0.0836	0.1283	0.2122	0.2515	0.0805	0.0408	0.0264	0.0428	0.2413
		LLM-Judge	0.5000	0.5000	0.5000	0.4000	0.5500	0.4500	0.5500	0.2500	0.1000	0.0000	0.0000	0.5500
	X2	EM	0.8000	0.7250	0.8250	0.2250	0.7000	0.6000	0.7000	0.3500	0.5000	0.2500	0.2500	0.7250
		BLEU-1	0.2543	0.2227	0.2147	0.1895	0.2091	0.3594	0.2910	0.2249	0.2182	0.0367	0.0843	0.2939
		LLM-Judge	0.8000	0.6000	0.8000	0.5500	0.4000	0.5500	0.6000	0.4000	0.2500	0.0000	0.0000	0.6000
	X3	EM	0.5500	0.5250	0.4750	0.2500	0.4000	0.4000	0.3000	0.3250	0.4500	0.3750	0.4750	0.2750
		BLEU-1	0.1523	0.0956	0.1541	0.0979	0.1053	0.1540	0.0111	0.0764	0.1096	0.1086	0.1054	0.0091
		LLM-Judge	0.3000	0.1000	0.2500	0.0000	0.3000	0.2500	0.0000	0.0500	0.2000	0.0500	0.0500	0.0000
	X4	EM	0.2583	0.1667	0.2083	0.2500	0.2250	0.1917	0.0250	0.1250	0.1833	0.2500	0.3417	0.0417
		BLEU-1	0.1967	0.1245	0.2085	0.1504	0.0603	0.0896	0.1191	0.0048	0.1757	0.0770	0.0413	0.1191
		LLM-Judge	0.1667	0.1333	0.2000	0.2167	0.0833	0.1167	0.1000	0.0333	0.2000	0.0667	0.1333	0.1000
Avg.	–	EM	0.5361	0.4950	0.5093	0.2486	0.4873	0.4885	0.3436	0.2974	0.3812	0.3295	0.3856	0.3434
		BLEU-1	0.2171	0.1775	0.1996	0.1415	0.1476	0.1831	0.1260	0.1249	0.1790	0.1040	0.0746	0.1260
		LLM-Judge	0.4017	0.3277	0.3639	0.2454	0.2741	0.3190	0.2390	0.2112	0.2518	0.1114	0.1021	0.2329

Table 19: Results on a subset of the MemEye evaluation matrix using `gemini-2.5-flash-lite`. Columns correspond to memory methods, grouped into text-only and multimodal families. Within each coordinate, EM is reported for multiple-choice questions, while BLEU-1 and LLM-as-a-Judge (LLM-Judge) are reported for free-response questions. The first- and second-performing memory model(s) are highlighted with orange and blue backgrounds, respectively.

Y	X	Metric	Textual memory							Multimodal memory				
			FC(T)	SRAG(T)	Refl.	Gen.Ag.	MemOS	A-Mem	SM(T)	FC(V)	SRAG(V)	MMA	M2A	SM(V)
Y1	X1	EM	0.7500	0.7250	0.6000	0.2500	0.4750	0.5750	0.1500	0.8750	0.9000	0.4000	0.3000	0.3750
		BLEU-1	0.2689	0.1777	0.2456	0.1278	0.1096	0.0672	0.2119	0.2841	0.3328	0.2005	0.0508	0.2099
		LLM-Judge	0.5500	0.4000	0.5556	0.2000	0.2000	0.4000	0.5500	0.6000	0.7500	0.3000	0.2000	0.4500
	X2	EM	0.6667	0.6111	0.5556	0.2500	0.4444	0.5278	0.1944	0.7500	0.5833	0.4167	0.0833	0.3611
		BLEU-1	0.1325	0.0613	0.0757	0.0807	0.1431	0.0870	0.0000	0.1192	0.1382	0.1013	0.0274	0.0190
		LLM-Judge	0.4444	0.2778	0.3333	0.1667	0.2222	0.2778	0.2222	0.6111	0.5000	0.1111	0.1111	0.2222
	X3	EM	0.4309	0.3914	0.2599	0.2533	0.2632	0.2763	0.0757	0.6776	0.7500	0.4868	0.1053	0.0855
		BLEU-1	0.2585	0.2203	0.1466	0.1061	0.0894	0.1035	0.0655	0.2928	0.3186	0.3552	0.0566	0.0293
		LLM-Judge	0.3553	0.2632	0.2039	0.1184	0.0658	0.1579	0.1382	0.4934	0.6053	0.5132	0.0724	0.0987
	X4	EM	0.4861	0.4583	0.3472	0.2639	0.4722	0.5000	0.0972	0.7917	0.8194	0.6667	0.1389	0.0972
		BLEU-1	0.2180	0.1982	0.1062	0.0205	0.0316	0.1875	0.0556	0.3300	0.2881	0.3139	0.0281	0.0556
		LLM-Judge	0.4167	0.3056	0.1667	0.0000	0.0556	0.1389	0.0556	0.6111	0.6389	0.5278	0.0278	0.0556
Y2	X1	EM	0.4397	0.4310	0.4741	0.2500	0.2759	0.3793	0.2328	0.5000	0.4655	0.3879	0.1638	0.2586
		BLEU-1	0.4360	0.3272	0.3461	0.1401	0.1279	0.1607	0.2682	0.3290	0.3624	0.4404	0.0100	0.3076
		LLM-Judge	0.5000	0.3793	0.3966	0.1724	0.1379	0.2069	0.3621	0.4138	0.4310	0.5345	0.0345	0.3966
	X2	EM	0.4295	0.3397	0.4295	0.2500	0.3397	0.3654	0.2500	0.4551	0.5256	0.3654	0.2051	0.2564
		BLEU-1	0.1546	0.1805	0.1795	0.1065	0.1540	0.1441	0.1001	0.1254	0.2260	0.0994	0.0516	0.1029
		LLM-Judge	0.2308	0.2051	0.2436	0.2051	0.2692	0.2308	0.1538	0.2564	0.3333	0.2308	0.0513	0.1538
	X3	EM	0.2857	0.2560	0.2738	0.2440	0.2738	0.2619	0.0536	0.3869	0.4048	0.2619	0.1548	0.0774
		BLEU-1	0.1576	0.1232	0.0838	0.0500	0.1460	0.1285	0.0096	0.1930	0.1762	0.1761	0.0225	0.0112
		LLM-Judge	0.2143	0.1071	0.0789	0.0952	0.1548	0.1667	0.0238	0.3810	0.3452	0.3571	0.0119	0.0238
	X4	EM	0.3210	0.2841	0.2727	0.2500	0.2585	0.2756	0.0511	0.4517	0.3778	0.2955	0.1534	0.0540
		BLEU-1	0.1449	0.1299	0.0660	0.0606	0.1315	0.0898	0.0357	0.2468	0.1323	0.1955	0.0327	0.0599
		LLM-Judge	0.1818	0.1648	0.1354	0.1023	0.1136	0.1193	0.0568	0.3011	0.2273	0.2443	0.0739	0.0795
Y3	X1	EM	0.7000	0.5500	0.6000	0.2500	0.3500	0.7250	0.1000	0.6250	0.4750	0.4750	0.1750	0.0750
		BLEU-1	0.1971	0.1808	0.2247	0.0987	0.1637	0.1656	0.1845	0.1279	0.1691	0.1685	0.0269	0.1634
		LLM-Judge	0.6000	0.4000	0.4375	0.3000	0.2500	0.4500	0.4500	0.4000	0.4000	0.4000	0.0500	0.4000
	X2	EM	0.8500	0.6500	0.8000	0.2500	0.0750	0.7750	0.1750	0.7500	0.5000	0.3750	0.1750	0.2000
		BLEU-1	0.2189	0.1631	0.2594	0.1656	0.1794	0.2737	0.1901	0.1237	0.1498	0.0956	0.0385	0.2227
		LLM-Judge	0.6500	0.5000	0.6000	0.6000	0.2000	0.6000	0.6000	0.5000	0.3000	0.0500	0.0000	0.5000
	X3	EM	0.5000	0.5750	0.3250	0.2750	0.4750	0.5250	0.0500	0.3750	0.5500	0.7000	0.4250	0.0500
		BLEU-1	0.1423	0.1601	0.0657	0.0310	0.0469	0.0709	0.0215	0.0669	0.0871	0.1157	0.0639	0.0015
		LLM-Judge	0.1000	0.2500	0.1500	0.0000	0.0000	0.0500	0.0000	0.0000	0.0000	0.1500	0.0000	0.0000
	X4	EM	0.2667	0.2333	0.2833	0.2417	0.2000	0.2500	0.0250	0.3000	0.2083	0.2750	0.0833	0.0500
		BLEU-1	0.0490	0.0808	0.0590	0.0148	0.0966	0.0943	0.0667	0.0470	0.1087	0.2124	0.0105	0.0667
		LLM-Judge	0.1333	0.1167	0.1154	0.0667	0.1000	0.2500	0.0667	0.0667	0.1500	0.2167	0.0000	0.0667
Avg.	–	EM	0.5105	0.4588	0.4351	0.2523	0.3252	0.4530	0.1212	0.5782	0.5467	0.4255	0.1802	0.1617
		BLEU-1	0.1982	0.1669	0.1549	0.0835	0.1183	0.1311	0.1008	0.1905	0.2074	0.2062	0.0350	0.1041
		LLM-Judge	0.3647	0.2808	0.2847	0.1689	0.1474	0.2540	0.2233	0.3862	0.3901	0.3030	0.0527	0.2039

E CASE STUDIES

Qualitative interpretation. The case studies illustrate the main failure modes measured by MemEye. Cases 1-5 focus on visual-evidence loss: replacing images with captions can remove decisive details even when the surrounding dialogue is preserved. Cases 6-11 focus on evolving visual states: retrieval can surface semantically related images, but the model must still determine which visual evidence is temporally valid. The final textual-memory examples show a complementary pattern: compact state abstractions can help track updates, but they must be paired with preservation of visual evidence.

Case 1: Cross-Session Identity Tracking

Cell: (X_3, Y_2)
FC-V = 1.00
FC-T = 0.00



S5: Pair by rocks | S5: Pair in valley | S9: Pair with egg

Q: A small green dinosaur and a brown bird appeared as a pair in two Episode 1 scenes. Were these same two characters also seen together holding a large egg in a later session?

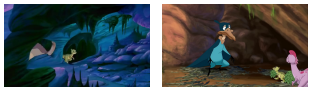
A: Yes—the same green dinosaur and brown bird appear consistently across all three scenes.

Captions seen by text-only model:
S5-R6: *"Two animated dinosaurs peek over rocks beside a stream in a canyon."*
S5-R7: *"Two cartoon dinosaurs stand in a rocky canyon beside a small stream."*
S9-R3: *"A green dinosaur stands on the back of a large yellow-spotted dinosaur in a rocky landscape."*

Why caption fails: The captions lose species identity, body shape, and skin texture—the visual cues needed to confirm the same pair across sessions.

Case 2: Micro-Attribute Comparison

Cell: (X_4, Y_2)
FC-V = 1.00
FC-T = 0.00



Ep. 1 cave (S1): stalactites | Ep. 2 cave (S7): vegetation

Q: Both episodes open with a cave scene. In Episode 1's dark cave, icicle-shaped rock formations hang from the ceiling. In Episode 2's cave, what hangs from the ceiling instead?

A: Green hanging vegetation and vines.

Captions seen by text-only model:
S1-R4: *"A small dinosaur stands inside a blue cave with stalactites and rocky walls."*
S7-R3: *"A blue bird stands in a cave facing two small dinosaurs."*

Why caption fails: S1's caption mentions "stalactites," but S7's caption omits the ceiling entirely, making the cross-episode comparison impossible.

Case 3: Fine-Grained Color Memory

Cell: (X_4, Y_2)
FC-V = 1.00
FC-T = 0.25



S5: First close-up | S10: Later close-up

Q: In the close-up where two characters sit on green grass—one purple, one orange—what are their exact eye colors?

A: Purple character: red eyes; orange character: green eyes.

Captions seen by text-only model:
S5-R3: *"A group of cartoon dinosaurs stands on green grass, with a small purple dinosaur holding an egg in the center."*
S10-R4: *"Two cartoon dinosaurs stand together outdoors."*

Why caption fails: Neither caption mentions eye color. The text-only model must guess, while the visual model reads the exact hue from the pixel data.

Figure 10: Three Caption-Proof examples from MemEye's Cartoon Ent. task. Each question requires native visual input to answer; replacing images with dense captions causes accuracy to collapse. **Case 1** tests cross-session character identity tracking (X_3, Y_2). **Case 2** tests micro-attribute comparison (X_4, Y_2). **Case 3** tests fine-grained color recall (X_4, Y_2). Together, these cases illustrate why MemEye's Caption-Proof protocol is essential: without it, benchmark scores may reflect text-based retrieval rather than genuine visual memory.

Case 1: Cross-session identity tracking. This case requires matching the same two characters across visually different scenes and sessions. The caption-only failure shows that object-category descriptions are insufficient; the model must preserve identity cues, such as body shape, color patterns, and character pairings, across time.

Case 2: Micro-attribute comparison. This question compares ceiling details across two cave scenes. Although the first caption mentions stalactites, the second omits the corresponding ceiling

Case 4: State-Evolving Belief Revision ($A \to B \to A$)

Task: Health Care *Cell:* (X_4, Y_3)

FC-V = 1.00 FC-T = 0.00



D3: "Pair carbs with protein"



D6: "Manage carb snacks"



D11: "Continue pairing carbs"

Q: Across all doctor portal messages Maya received during the month, what is the focus of the MOST RECENT guidance from Dr. Ramirez?

A: Continue pairing carbs with protein or fiber.

Captions seen by text-only model:
D3: *"A health portal message from a doctor advises the patient to keep pairing carbs with protein or fiber. . . "*
D6: *"A health portal message shows a doctor advising a patient on managing small carb snacks around workouts."*
D11: *"A patient views a follow-up nutrition message from their doctor in the Northline Health Portal."*

Why caption fails: D3 and D6 captions reveal their content, but the critical D11 caption is generic and does not preserve which guidance remains current. A text-only model sees that a follow-up exists but lacks the D11 state needed to complete the $A \to B \to A$ chain, so it answers B.

Case 5: Game-State Tracking Under Updates

Task: Card Playlog *Cell:* (X_4, Y_3)

FC-V = 1.00 FC-T = 0.25



UNO state after Player 3's hand changes from 5 to 4 cards

Q: Immediately after Player 3's visible hand size changes from 5 to 4 for the 1st time, how many red cards does Player 2 hold?

A: 3

Caption seen by text-only model:
"Digital UNO game board showing four AI players, a central pile of played cards with a blue Reverse as the target card, a. . . "

Why caption fails: The caption describes the board layout generically but omits per-player hand composition. Counting red cards in a specific player's hand at a precise game state requires reading the actual card faces from the screenshot—a fine-grained visual task (X_4) combined with temporal state tracking (Y_3).

Figure 11: Two additional Caption-Proof cases highlighting MemEye's Y_3 state-evolving synthesis dimension. **Case 4** demonstrates an $A \to B \to A$ belief reversal across three doctor portal messages: the model must preserve the most recent portal-message state to discover that the final guidance reverts to the initial advice, a fact invisible in the generic D11 caption. **Case 5** requires counting specific card colors in a player's hand at a precise temporal state—information that captions never enumerate. Both cases show that MemEye's hardest region, high $Y \times$ high X, demands joint visual fidelity and temporal reasoning that text surrogates cannot support.

feature. As a result, the text-only model lacks the contrastive evidence needed to infer that the later cave contains hanging vegetation.

Case 3: Fine-grained color memory. This case isolates a pixel-level attribute that captions often discard: exact eye color. The visual model can inspect the characters directly, while the captioned representation collapses the relevant cue into a generic character description.

Case 4: State-evolving belief revision. The doctor-message example tests whether a model can track an $A \to B \to A$ update chain. The final message is visually decisive because it restores the original recommendation, but its caption is too generic. This shows that state revision can fail when the latest evidence is available only through native visual text.

Case 6: Silent Tag Override
Task: CrossScene Memory
Cell: (X_4, Y_3)
FC-V = 0.75 RAG-V = 0.00



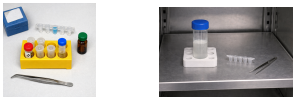
S7: Tag reads "C-1127" (×3 images) | S9: Tag reads "A-209" (×1 image)

Q: What identification tag number is currently displayed in the fossil room case?

A: A-209

Why RAG fails:
Retrieval for "fossil room tag" returns 4 images: 3 show "C-1127" (S7) and 1 shows "A-209" (S9). A frequency-voting retrieval system answers "C-1127" (3:1 majority). Only temporal reasoning—understanding that S9 is **later** than S7—yields the correct **current** tag "A-209."

Case 7: Object Migration
Task: CrossScene Memory
Cell: (X_2, Y_3)
FC-V = 1.00 RAG-V = 0.25



S10: Forceps on main bench | S11: Forceps on cold-room shelf

Q: Where are the bent-tip forceps that were originally on the main lab bench?

A: On the lower shelf of the cold-room prep area.

Why RAG fails:
Retrieval for "bent-tip forceps" returns images from both the original location (S10, main bench) and the current location (S11, cold-room shelf). Without temporal ordering, RAG cannot determine which is the **current** position. The question asks "where are they," not "where were they."

Case 8: Narrative Arc Tracking
Task: Cartoon Ent. (Comic)
Cell: (X_1, Y_3)
FC-T = 0.75 RAG-T = 0.25



P28: Pelted by crowd | P29: Forced to do chores | P31: Angry outburst

Q: I feel like the temporary ruler's situation actually got better over time. Did it really get worse or better?

A: It got worse.

Why RAG fails:
Individual pages show mixed signals: P28 shows pelting *then* relaxing, P29 shows chores, and P31 shows a steak dinner. A retrieval system may surface the "relaxing" or "steak" fragments and conclude improvement. Only tracking the **full temporal arc** across pages reveals a consistent downward trend.

Figure 12: Three Y_3 cases where retrieval-based methods systematically fail and temporal reasoning is required. **Case 6**: old evidence outnumbers new evidence 3:1—frequency voting picks the stale tag. **Case 7**: the same object appears at two locations across sessions—only temporal ordering identifies the current one. **Case 8**: the narrative arc contains local ups and downs—only tracking the full trajectory reveals the overall decline. These cases illustrate why MemEye's Y_3 dimension is necessary: it separates models that can maintain a coherent, temporally ordered world model from those that treat memory as a static bag of retrieved fragments.

Case 5: Game-state tracking under updates. The UNO example combines temporal triggering with fine-grained visual counting. The answer depends on identifying the first-hand-size transition and then counting another player's red cards at that exact state. A generic board caption does not preserve this structured visual evidence.

Case 6: Silent tag override. The fossil-tag example illustrates a stale-majority failure. Retrieval surfaces more old images than new images, so a frequency- or similarity-based strategy favors the

outdated tag. The correct answer requires recognizing that the later visual state overrides the earlier one.

Case 7: Object migration. The forceps example requires tracking an object as it moves from one location to another. Both locations are semantically relevant, but only temporal ordering distinguishes the historical location from the current one. This exposes the gap between evidence retrieval and state resolution.

Case 8: Narrative arc tracking. The comic example shows that local retrieval snippets can be misleading when the answer depends on a trajectory. Individual pages contain both positive and negative events, but the correct interpretation requires integrating the sequence and judging the overall direction of change.

Case 9: Object reappears nearby. This visual-state probe shows a retrieval miss rather than a recognition failure. The backpack is absent from the old desk close-up but visible nearby in a later scene. When retrieval misses the later evidence, the model cannot recover the correct state even though the visual cue is unambiguous.

Case 10: Card-state trigger missed. This case shows that both the trigger event and the answer state are necessary. Retrieving only the later card screenshot is insufficient unless the system also identifies the preceding hand-size change that defines the relevant moment.

Case 11: Plastic-bag stale trap. The plastic-bag example shows how a visually salient old state can dominate retrieval. All three methods retrieve the earlier shelf evidence but miss the later desk evidence near the keyring, leading them to answer with the stale location rather than the updated state.

Case 9: Object Reappears Nearby

Q: Which object changed from not visible in the later desk close-up to present elsewhere nearby? **A:** The black backpack with the silver star charm.

	E1: old desk, absent	E2: latest doorway	Diagnosis
Oracle	 old desk	 backpack remains	Full chain.
SRAG(V)	 retrieved	missing	Misses backpack updates; retrieves cabinet evidence and answers red USB.
MMA	missing	missing	Misses backpack chain; retrieves off-target desk evidence.
M2A	missing	missing	No backpack evidence retrieved.

Case 10: Card State Trigger Missed

Q: How many red cards does Player 3 hold immediately after Player 2's visible hand size changes from 4 to 5 for the first time? **A:** 3.

	E1: trigger state	E2: answer state	Diagnosis
Oracle	 Player 2: 4 cards	 Player 2: 5 cards	Compare before/after.
SRAG(V)	missing	 retrieved	Misses trigger; retrieves old game state and gives wrong count.
MMA	missing	 retrieved	Misses trigger; retrieves old game state and gives wrong count.
M2A	missing	answer state missing	No needed game state.

Continued on next page.

Case 11: Plastic Bag Stale Trap

Q: Which scenario describes the transparent plastic bag? **A:** It was on the cabinet's lower shelf earlier and later appeared on the desk near the keyring.

	E1: old shelf	E2: latest desk	Diagnosis
Oracle	 bag on shelf	 bag near keyring	Shelf → desk.
SRAG(V)	 retrieved	missing	Retrieves old shelf close-up; answers stale shelf state.
MMA	 retrieved	missing	Retrieves old shelf close-up; answers stale shelf state.
M2A	 retrieved	missing	Misses latest desk image.

Figure 13: Retrieval-error case studies from the Y_3 visual-state update probe. Each panel aligns the oracle evidence chain with the answer-relevant evidence retrieved by SRAG(V), MMA, and M2A. **Case 9** shows a nearby-object reappearance that retrieval misses. **Case 10** shows a trigger-state miss in a card-game update. **Case 11** shows a stale-evidence trap where the transparent plastic bag is retrieved from its earlier shelf location, while the later desk state is missing. These examples isolate retrieval-side failures: the methods miss the decisive updated or comparison state and instead supply stale or off-target visual evidence. Irrelevant top-k distractors are omitted for readability.

E.1 Textual Memory Case Study

Table E.1 presents two Y_3 case studies where text-based memory preserves the evolving state chain while multimodal memory retrieves stale, conflicting, or visually similar evidence. These cases suggest that the advantage of A-Mem in some Y_3 scenarios comes from compact state extraction rather than richer visual recall: when the answer depends on tracking how visual states evolve, a structured textual evidence chain can be more reliable than retrieving visually similar raw images.

Brass compass migration *Question:* Where is the brass compass now? *Ground truth:* restoration table
Evidence chain: fossil case with brass compass → compass missing from fossil case → same brass compass appears on restoration table.

Method	Retrieved textual state / evidence chain	Retrieved image evidence	Answer
A-Mem	**Retrieved textual states:** (1) fossil display case with a "brass compass"; (2) the left-side metal object "is still missing" from the fossil case; (3) the restoration table contains a "vintage brass compass" beside the vial and tool.	/	**Correct.** Uses the updated restoration-table state.
M2A	Top semantic memories focus on "left-side absence remains" and nearby off-task brass/prop evidence. The current restoration-table update is not surfaced as the merged state.	 stale fossil absence; off-task brass object; prop watch distractor	**Wrong.** Says the compass is missing or points to an off-target table.
MMA	Retrieves raw fossil-room image turns and visually similar later shots, but no explicit "compass moved to restoration table" state.	 old fossil case; missing state; later fossil room	**Wrong.** Keeps a stale/off-target location.

Sage-green paint rejected *Question:* Which tested paint color was not used? *Ground truth:* sage green
Evidence chain: sage-green paint test → design pivots to terracotta → final room uses terracotta, so sage green is the tested color not used.

Method	Retrieved textual state / evidence chain	Retrieved image evidence	Answer
A-Mem	**Retrieved textual states:** (1) a "sage green paint" swatch is tested on the wall; (2) a later wall test uses "terracotta paint"; (3) the final living room has a "terracotta accent wall."	/	**Correct.** Names sage green as tested but rejected.
M2A	Retrieves the green test and final-room evidence, but abstracts the color as "muted olive-green" or generic "green wall paint," losing the exact label.	 terracotta test; green test; final room	**Wrong.** Loses the exact sage-green answer.
MMA	Retrieves separate raw visual entries for terracotta testing, the earlier green test, and the final room, but does not extract "tested but not final."	 terracotta test; sage test; final room	**Wrong.** Selects terracotta, the final color, rather than the rejected test color.

Figure 14: Case study of textual memory extraction and image retrieval for two Y_3 evolving-state cases. Each panel first specifies the required evidence chain, then compares the evidence retrieved by A-Mem, M2A, and MMA at answer time.