MEMCANVAS: VISUAL LONG-TERM MEMORY FOR MLLM-BASED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid advances in multimodal large language models (MLLMs), agents built upon them remain fundamentally stateless, processing each interaction in isolation with no capacity to retain or reuse experiential knowledge. Existing lifelong memory systems, whether parametric or retrieval-based, operate predominantly in text, discarding the rich visual structure of images, tables, and diagrams or reducing them to lossy textual descriptions. We present **MemCanvas**, a training-free, model-agnostic visual memory framework that stores heterogeneous multimodal experiences as structured canvas images, allowing agents to accumulate and retrieve knowledge through their native visual perception channel. MemCanvas introduces three coordinated mechanisms: an adaptive layout algorithm that compresses and assembles multimodal elements into structured visual canvases, a hybrid retrieval system that fuses visual and textual encoder embeddings for cross-modal memory search, and a frequency-adaptive consolidation strategy that progressively degrades rarely-accessed memories to curb storage growth. Extensive experiments across five benchmarks with two backbones show that MemCanvas improves direct inference by $+17.5$ EM on HotpotQA, outperforms text-based memory baselines (Mem0, MemVerse, A-Mem, Text-RAG) without task-specific training—except on purely text-only questions, where text stores remain preferable—and matches the best classical eviction strategy per retained memory at matched retention while degrading memories progressively instead of deleting them.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (OpenAI, 2023; Bai et al., 2025; Chen et al., 2024) have achieved remarkable capabilities across vision–language tasks. Yet agents built upon these models remain fundamentally stateless: each query is processed in isolation, with no mechanism to accumulate, retain, or reuse knowledge from prior interactions. Equipping MLLM-based agents with lifelong memory—the ability to persistently store, retrieve, and consolidate experiential knowledge—has therefore emerged as a critical research challenge (Zhang et al., 2025).

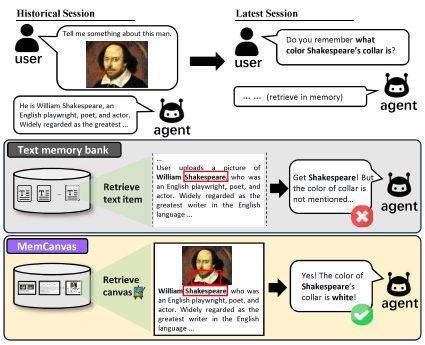

Figure 1: While existing text-based memory systems suffer from significant detail loss when handling multimodal information, MemCanvas renders that onto a canvas image as memory content, enabling the MLLM to read it through its natural visual channel.

Existing lifelong memory systems fall into three paradigms, all predominantly text-centric: parametric (e.g., WISE (Wang et al., 2024), MemLoRA (Bini et al., 2025)), non-parametric text stores (e.g., MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025)), and hybrid text knowledge graphs (e.g., MemVerse (Liu et al., 2025)). None of these preserves original visual structure: images, tables, and diagrams are discarded or reduced to textual descriptions. A critical gap remains: **no existing system provides visual memory over heterogeneous multimodal evidence—native images, tables, and**

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

Table 1: Comparison with lifelong memory systems for LLM-based agents. Type: P = parametric, NP = non-parametric, H = hybrid. MemCanvas is the only listed non-parametric system that stores multimodal evidence in a unified visual format without task-specific training.

Method	Type	Train-Free	Multimodal	Mem. Format	Consolidation
MemGPT	NP	✓	✗	Text	✗
Mem0	NP	✓	✗	Text	✓
A-Mem	NP	✓	✗	Text	✗
WISE	P	✗	✗	Params	✗
MemVerse	H	✗	✓	KG	✓
VizoMem (Liang et al., 2026)	NP	✗	✗	Image	✓
V-Mem (Kang et al., 2026)	NP	✓	✓	Raw units	✗
CMA (Wang et al., 2026a)	P	✗	✓	Thumbnails+text	✗
MemCanvas	**NP**	✓	✓	**Image**	✓

charts—that is non-parametric, training-free, and equipped with automatic consolidation. As illustrated in Figure 1, when a user shows a portrait of Shakespeare and later asks about the collar color, text-based memory systems that discard visual information cannot answer, whereas visual memory that preserves the original image can.

We address this gap with a key observation: MLLMs are inherently skilled at extracting structured information from images, so memories presented as images can be absorbed through the native visual perception channel, bypassing the text-token bottleneck. Recent optical compression methods (Shi et al., 2026; Feng et al., 2026) optimize the token cost of rendering a single given context; MemCanvas instead targets a persistent memory bank in which heterogeneous multimodal episodes are stored as canvases, retrieved cross-modally, and consolidated over time. These insights motivate **MemCanvas**, a non-parametric visual long-term memory framework that renders heterogeneous multimodal experiences as structured canvas images. A lightweight LLM compressor first condenses verbose content, and multimodal elements are then assembled into canvas images via a layout algorithm. Each canvas is indexed by hybrid dual-encoder embeddings that enable cross-modal similarity search. To curb memory growth, a frequency-adaptive consolidation mechanism progressively degrades low-frequency memories through resolution reduction.

Realizing this framework poses three challenges: unifying heterogeneous formats into one visual representation without information loss (adaptive layout), bridging modalities in retrieval (hybrid dual-encoder embeddings), and principled consolidation of a growing bank (frequency-adaptive forgetting).

Our contributions are as follows:

- We propose **MemCanvas**, a visual long-term memory system for MLLM-based agents that combines non-parametric storage of heterogeneous multimodal experiences as structured canvas images with hybrid cross-modal retrieval and automatic consolidation.
- We develop a visual-memory pipeline that integrates adaptive canvas construction, hybrid visual–text retrieval, and frequency-adaptive forgetting, enabling compact storage, cross-modal access, and automatic memory consolidation without task-specific training.
- We validate MemCanvas across five benchmarks spanning visual QA, multi-hop reasoning, chart understanding, and multimodal knowledge retrieval, showing gains for both open-source and proprietary MLLMs, with the largest improvements on knowledge- and visually-intensive questions.

2 Related Work

2.1 Memory for LLM Agents

Parametric memory encodes knowledge directly into model parameters. WISE (Wang et al., 2024) proposes a dual parametric memory scheme with main and side memories for lifelong knowledge editing, and MemLoRA (Bini et al., 2025) distills expert adapters for on-device memory systems.

While parametric approaches integrate seamlessly with inference, they require retraining for each new piece of knowledge and struggle with catastrophic forgetting.

Non-parametric memory maintains external storage structures accessed during inference. MemGPT (Packer et al., 2023) pioneers virtual context management with hierarchical memory tiers. A-Mem (Xu et al., 2025) introduces Zettelkasten-inspired agentic memory with dynamic indexing and linking. Mem0 (Chhikara et al., 2025) provides a scalable memory layer that dynamically extracts and consolidates salient information from multi-session dialogues.

Multimodal memory relaxes the text-only assumption. M3-Agent (Long et al., 2025), WorldMM (Yeo et al., 2025), and AtlasVA (Wang et al., 2026b) add multimodal episodic and semantic tiers for long-video and embodied settings but rely on RL training; ViLoMem (Bo et al., 2025) grows visual guidelines from solving errors rather than storing episodes. V-Mem (Kang et al., 2026) routes retrieval by query modality—a query-adaptive alternative to our fixed mixing ratio—CMA (Wang et al., 2026a) retrieves episodic thumbnails with a trained controller, MMA (Lu et al., 2026) reweights memories by credibility and recency, and VimRAG (Wang et al., 2026c) organizes visual context into a trained memory graph. On consolidation, SimpleMem (Liu et al., 2026b) shows compressed multi-view text units suffice for text-centric agents, and LUMA-RAG (Wandre et al., 2025) adds provably stable streaming cross-modal alignment. Mem-Gallery (Bei et al., 2026), MemEye (Guo et al., 2026), DMV-Bench (Tang et al., 2026), and WorldMemArena (Liu et al., 2026a) provide benchmarks probing multimodal long-term memory and fine-grained visual recall. MemCanvas differs from these systems in storing heterogeneous episodes themselves—questions, answers, images, and tables—as unified visual artifacts with training-free consolidation.

Despite these advances, a critical limitation persists: *nearly all existing memory systems operate exclusively in text*. MemVerse (Liu et al., 2025) accepts multimodal input but maps visual content to textual knowledge-graph nodes, losing original visual structure. Table 1 positions MemCanvas as the only listed non-parametric framework that supports multimodal memory in a unified visual format with automatic consolidation, while requiring no task-specific training.

2.2 Optical Compression

A recent line of work demonstrates that MLLMs can effectively read text rendered as images: DeepSeek-OCR (Wei et al., 2025; 2026) renders long text contexts into 2D images with $\sim$97% OCR precision at up to $\sim$10$\times$ compression, Glyph (Cheng et al., 2025) reports 3–4$\times$ compression with accuracy on par with text baselines, and Text or Pixels? (Li et al., 2025) retains 97–99% accuracy with nearly 50% fewer tokens; AgentOCR (Feng et al., 2026) and MemOCR (Shi et al., 2026) extend the paradigm to agent histories and rich-text memory. Closest to our setting, VizoMem (Liang et al., 2026) renders text into structured "visual notes" retrieved by a purpose-trained late-interaction encoder, showing that image-based memory units can match text stores at a fraction of the token cost. VizoMem, however, is designed for text histories and trains its retrieval encoder: it does not construct canvases from native multimodal evidence, combine off-the-shelf image and text embeddings for cross-modal retrieval, or degrade resolution for consolidation—the three mechanisms whose combination distinguishes MemCanvas. These works validate a key insight: *the visual modality is not merely an input channel but a viable information carrier*, so heterogeneous multimodal information can be *natively unified* into a single visual format rather than compressed text alone.

3 Method

3.1 Overview

As illustrated in Figure 2, MemCanvas consists of three modules. **Memory Construction** converts multimodal data—text, images, and tables—into structured canvas images readable through the MLLM's native visual channel. **Storage & Retrieval** indexes each memory with dual embeddings (one text, one image) for cross-modal retrieval. **Update & Forgetting** progressively degrades rarely accessed memories to keep the store compact. At query time, the query is encoded by a text encoder, the top-K most similar canvases are retrieved by hybrid cosine similarity, and they are injected as visual context alongside the original query.

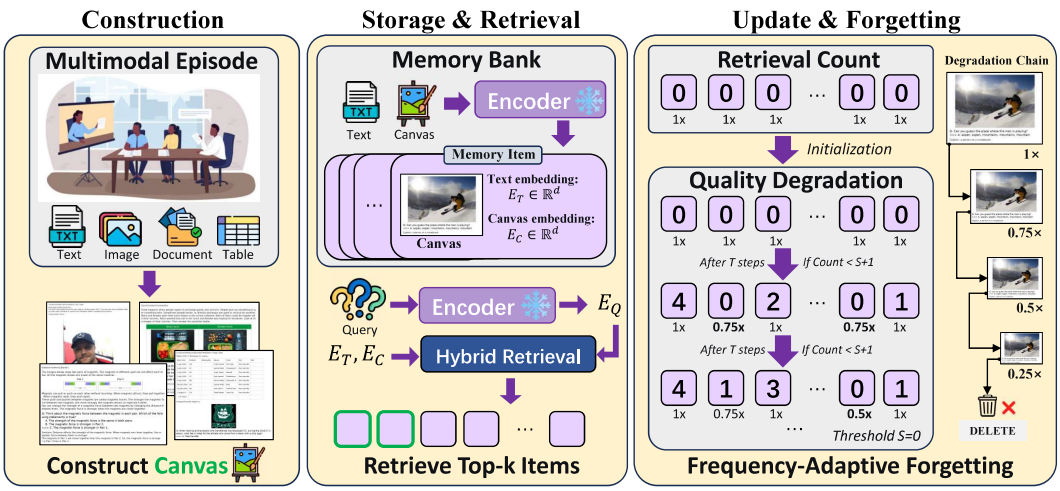

Figure 2: Overview of MemCanvas. Memory Construction converts multimodal conversations into structured canvas images via preprocessing, rectangularization, and adaptive layout optimization; Storage & Retrieval indexes canvases with hybrid dual-encoder embeddings for efficient cross-modal search; Update & Forgetting maintains a compact memory bank through frequency-adaptive resolution degradation.

3.2 Memory Construction

The Memory Construction pipeline transforms multimodal conversational records into single, unified canvas images through four stages (Figure 3).

3.2.1 Stage 1: Preprocessing

Each modality undergoes format-specific preprocessing to produce visual-ready content blocks. Text is compressed by a lightweight LLM (Qwen3-4B) into a compact form without task-specific training. Images, tables, and documents are resized while maintaining their aspect ratios, with column widths and row formatting optimized for readability.

3.2.2 Stage 2: Rectangularization

After preprocessing, each visual block is classified into one of two rectangle types. *Text blocks* are **flexible**: text can reflow across multiple lines at varying widths. *Image blocks* are **rigid**: images have a fixed aspect ratio, and tables have fixed row heights but flexible column widths.

3.2.3 Stage 3: Candidate Generation

The layout strategy follows a Measure-First, Build-Later approach: after estimating the total content area, the strategy derives a target canvas width and generates candidate layouts from predefined styles—single-column stacking, and two-column layouts with full-width images and side-by-side text or tables. *Text reflow* lets the same content produce different canvas shapes at different widths, enabling the optimizer to pack the canvas efficiently while maintaining clarity.

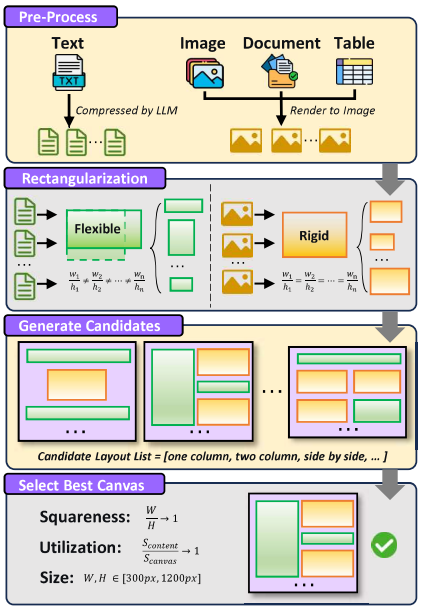

Figure 3: Adaptive canvas layout pipeline. Preprocessing converts each modality into visual-ready blocks, rectangularization characterizes their dimensional flexibility, candidate generation produces multiple layouts at varying widths, and the best canvas is selected by a multi-objective scoring function.

3.2.4 STAGE 4: BEST CANVAS SELECTION

The best layout is selected by a multi-objective scoring function:

$$S(l) = 0.6 \cdot \underbrace{\min\left(\tfrac{W}{H}, \tfrac{H}{W}\right)}_{\text{squareness}} + 0.3 \cdot \underbrace{\frac{\sum_i a_i}{W \cdot H}}_{\text{utilization}} + 0.1 \cdot \underbrace{\mathbb{I}[W, H \in \text{bounds}]}_{\text{size}}, \tag{1}$$

where *squareness* penalizes elongated canvases (encouraging aspect ratios close to 1:1 for efficient vision-encoder tiling), *utilization* measures the fraction of canvas area occupied by content, and the *size* term ensures dimensions remain within acceptable bounds. The weights encode this priority ordering: squareness governs vision-encoder tiling cost and dominates the score, utilization proxies information density, and the size term acts as a feasibility constraint rather than a graded objective—when all candidates satisfy the bounds, selection falls back to the weighted combination of squareness and utilization. Candidates are generated by sweeping the canvas width over a discrete grid from 800 to 1600 px in 100 px steps and instantiating three predefined styles—single-column, two-column, and side-by-side—with text reflow evaluated at each width; all candidates are rendered at a fixed resolution with a 16 px body font, 20 px headings, 24 px inter-block padding, and images scaled to a maximum side of 600 px, so the scoring stage compares like-for-like. The size bounds keep the canvas within $W \in [800, 1600]$ and $H \in [400, 2000]$ pixels, so each canvas stays within a small, fixed number of vision-encoder tiles of the downstream MLLM. When several candidates tie, the one with the smallest total area is selected. The highest-scoring candidate is rendered as the final canvas image.

3.3 MEMORY STORAGE AND RETRIEVAL

Storage. Each memory entry in the bank consists of three components: (1) the rendered canvas image $\mathbf{c}_i$; (2) a text embedding $\mathbf{E}_i^{\mathrm{T}} \in \mathbb{R}^d$ produced by a text encoder from the original question, answer, and key metadata; and (3) an image embedding $\mathbf{E}_i^{\mathrm{C}} \in \mathbb{R}^d$ produced by an image encoder from the canvas image. We extract text embeddings in addition to image embeddings because the question, answer, and metadata carry dense semantic signals that pixel-level representations may underweight; the two channels capture orthogonal aspects of the same memory, and their fusion retrieves more robustly than either alone.

Retrieval. At inference, we construct a hybrid key embedding for each memory by fusing the two representations:

$$\mathbf{E}_i^{\mathrm{K}}(\alpha) = \mathrm{norm}\left(\alpha \cdot \mathbf{E}_i^{\mathrm{C}} + (1-\alpha) \cdot \mathbf{E}_i^{\mathrm{T}}\right), \tag{2}$$

where $\alpha \in [0, 1]$ controls the image–text mixing ratio. Both encoders output L2-normalized embeddings, so $\mathbf{E}_i^{\mathrm{C}}$ and $\mathbf{E}_i^{\mathrm{T}}$ are unit vectors before fusion. The text input concatenates the question, the gold answer, and the metadata fields (subject, grade, and topic for ScienceQA; question modality for MultiModalQA) in that order. Because queries are text-only, retrieval reduces to a text query matched against a statically fused key; early fusion keeps one index and one similarity computation per memory. Table 6 selects $\alpha = 0.75$; query-adaptive fusion is a natural extension. A new query q is encoded through the text encoder as $\mathbf{E_q}$, and the top-K most similar canvases are retrieved by cosine similarity:

$$\mathcal{M}_q = \text{top-}K\left(\left\{\frac{\mathbf{E_q} \cdot \mathbf{E}_i^{\mathrm{K}}(\alpha)}{\|\mathbf{E_q}\| \, \|\mathbf{E}_i^{\mathrm{K}}(\alpha)\|}\right\}_{i=1}^{N}\right). \tag{3}$$

The retrieved canvases $\mathcal{M}_q$ are injected as visual context alongside the query.

3.4 MEMORY UPDATE AND FORGETTING

As the bank grows, we propose a frequency-adaptive forgetting mechanism inspired by human memory decay. Every T queries, we inspect the cumulative retrieval count of each memory. Memories with count $\leq S$ move one state down a five-level quality ladder:

$$s_0 \xrightarrow{\text{degrade}} s_1\,(0.75\times) \to s_2\,(0.5\times) \to s_3\,(0.25\times) \to s_4\,(\text{delete}). \tag{4}$$

The $0.75\times$ factors scale width and height linearly (lossless PNG storage, so bytes shrink roughly quadratically), and $\mathbf{E}_i^{\mathrm{C}}$ is recomputed from the degraded canvas at each stage so the index always

describes the retrievable artifact ($\sim$34 ms/entry/stage). Retrieval counters are *cumulative* (never reset), so a memory must fail the threshold check at four consecutive checkpoints before deletion; any memory retrieved more than S times in total is protected thereafter, though its resolution is never restored. The policy is a frequency heuristic rather than a hard byte budget, and a strict bound can be layered on by capping entries kept at full resolution.

A key advantage of canvas memory is *continuous, graceful degradation*. Text stores can also be compressed—by progressive summarization, truncation, or field pruning—but every compression step requires deciding which content to verbalize or discard. Image resolution degradation requires no such decisions: reducing a canvas to $0.75\times$ preserves global layout, headings, and key figures while substantially reducing storage, and even at $0.5\times$ most textual content remains legible. MemCanvas thus turns forgetting from a discrete keep-or-discard decision into a continuous storage–fidelity trade-off.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on five QA benchmarks spanning visual QA, cross-modal reasoning, chart understanding, and multi-hop text reasoning: **ScienceQA** (Lu et al., 2022) (12,726 train / 4,241 test; accuracy), **OK-VQA** (Marino et al., 2019) (9,009 / 5,046; accuracy), **MultiModalQA** (Talmor et al., 2021) (23,817 / 2,441; EM/F1), **HotpotQA** (Yang et al., 2018) (a 50,000-example training subset / 7,405-example dev set; EM/F1), and **ChartQA** (Masry et al., 2022) (28,299 train / 1,920 validation; accuracy).

Table 2: HotpotQA results. Multi-hop reasoning over documents (7,405 dev, distractor setting). Text-RAG, Mem0, and the Qwen2.5-VL-7B rows below the midrule share the Qwen2.5-VL-7B backbone and our retrieval protocol.

Method	EM	F1
Baseline	45.60	59.02
Self-RAG (7B)	40.20	54.30
Self-RAG (13B)	44.20	58.20
IRCoT (GPT-3)‡	49.30	60.70
EfficientRAG	50.59	57.93
DSP‡	51.40	62.90
PRISM	54.20	66.96
Open-RAG (7B)	**63.30**	**76.90**
Text-RAG	44.36	60.15
Mem0	47.35	62.24
Qwen2.5-VL-7B	33.24	43.44
Qwen2.5-VL-7B + MemCanvas	50.70	65.72
GPT-4o-mini	60.20	73.44
GPT-4o-mini + MemCanvas	<u>60.80</u>	<u>75.03</u>

Protocol. Following the memory-bank protocol of prior memory systems (Chhikara et al., 2025; Xu et al., 2025; Liu et al., 2025), the training split of each benchmark is converted into canvas memories, forming the memory bank; the evaluation split is then answered query-by-query with the top-K retrieved canvases injected as visual context. Each canvas packages the training example's question, gold answer, and key metadata together with its associated images or tables. Because the bank is built from labeled training examples, the protocol measures retrieval-augmented inference with stored demonstrations rather than online experience accumulation; it isolates the effect of memory representation and retrieval, while Appendix A.3 reports results on the long-term conversational benchmark LoCoMo and Appendix G reports an answer-removal ablation and format-isolation comparisons. All controlled comparisons are within-model: the same backbone answers the same questions under each condition, with identical prompt templates and greedy decoding.

Baselines. We compare against (i) *direct inference* without memory; (ii) retrieval-format variants—*Random ICL* (randomly sampled training examples as text), *Text-RAG* (retrieved examples as text passages), and *Image Fewshot* (retrieved original training images)—which use the same training records as MemCanvas and, for the retrieved variants, the same CLIP retrieval and identical K; and (iii) memory-system baselines—MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025), WISE (Wang et al., 2024), and MemVerse (Liu et al., 2025). VizoMem (Liang et al., 2026) is not run head-to-head because its ColGlyph encoder is purpose-trained with no off-the-shelf checkpoint; Text-RAG is the controlled text-retrieval counterpart. For HotpotQA we additionally report published retrieval-augmented systems—Self-RAG (Asai et al., 2024), IRCoT (Trivedi et al., 2023), EfficientRAG (Zhuang et al., 2024), DSP (Khattab et al., 2022), PRISM (Nahid & Rafiei, 2025), and Open-RAG (Islam et al., 2024)—and for ScienceQA prior published results: Human, GPT-4, and CoT (Lu et al., 2022; Liu et al., 2023), DDCoT (Zheng et al.,

Table 3: Performance comparison (%) on ScienceQA. Categories: NAT (natural science), SOC (social science), LAN (language), TXT (text context), IMG (image caption), NO (no context), G1-6 (grades 1–6), G7-12 (grades 7–12). Rows above the second midrule and the MemVerse rows are from prior work (Lu et al., 2022; Liu et al., 2023; Zheng et al., 2023; Yao et al., 2023; Liu et al., 2025); the remaining rows are our runs. Best in **bold**, second best underlined.

	Subject			Context Modality			Grade		
Model + Method	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	**Avg**
Human	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-4	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
CoT (GPT-4)	85.48	72.44	**90.27**	82.65	71.49	**92.89**	86.66	79.04	83.99
DDCoT	80.15	76.72	82.82	78.89	72.53	85.02	82.86	75.21	80.15
HoT-T5-Large	84.46	79.08	84.64	82.89	75.81	88.15	83.88	82.47	83.38
Qwen2.5-7B	76.20	67.83	77.27	74.49	65.79	79.02	77.72	69.35	74.72
Qwen2.5-7B + MemVerse	74.51	68.50	78.73	75.92	66.19	81.95	79.70	64.73	75.62
GPT-4o-mini + MemVerse	85.26	81.55	89.09	83.28	78.19	91.50	88.11	80.75	85.48
Qwen2.5-VL-7B + A-Mem	89.75	90.10	86.80	88.90	**91.20**	88.40	91.10	85.55	89.03
Qwen2.5-VL-7B + Mem0	90.15	89.95	86.20	93.95	88.25	88.15	91.35	85.90	89.13
Qwen2.5-VL-7B	88.35	88.10	86.90	92.40	86.20	87.10	89.60	84.80	87.90
Qwen2.5-VL-7B + MemCanvas	**90.28**	90.55	88.00	95.06	88.05	89.20	91.19	**87.15**	**89.74**
GPT-4o-mini	78.40	82.50	76.10	84.20	77.30	78.10	81.20	74.60	79.00
GPT-4o-mini + MemCanvas	89.06	**94.17**	86.36	**95.92**	87.76	87.88	**92.21**	84.36	89.40

2023), and HoT-T5 (Yao et al., 2023). Storage comparisons in the appendix additionally cover the parametric systems WISE (Wang et al., 2024), MemLoRA (Bini et al., 2025), and MemOCR (Shi et al., 2026).

Models and implementation. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and GPT-4o-mini (OpenAI, 2024) as the primary MLLMs. The retrieval encoder is CLIP ViT-L/14 (Radford et al., 2021), producing 768-dimensional embeddings for both images and text; the text compressor is Qwen3-4B. Default retrieval parameters: $\alpha = 0.75$, top-$K = 2$, $T = 1000$, $S = 0$. A single global configuration of all hyperparameters—retrieval settings, layout weights, and the encoder—is applied identically across all five benchmarks; no benchmark-specific tuning is performed. All experiments run on a single NVIDIA A6000 GPU.

4.2 Main Results

HotpotQA. Table 2 shows that MemCanvas substantially improves the Qwen2.5-VL-7B baseline (EM 33.24 $\rightarrow$ 50.70, F1 43.44 $\rightarrow$ 65.72), outperforming Mem0 ($+3.35$ EM) and Text-RAG ($+6.34$ EM) without task-specific training, and also improves GPT-4o-mini (EM 60.20 $\rightarrow$ 60.80), indicating that the memory design transfers to stronger closed-source models. Published systems above the midrule use heterogeneous pipelines and serve as context; controlled conclusions rest on the within-model rows sharing one backbone and retrieval protocol. With a 7B VLM, MemCanvas is competitive with Wikipedia-search RAG systems (DSP 51.4, EfficientRAG 50.59, PRISM 54.20) despite their larger text-only LLMs; Open-RAG (63.3) remains ahead via a retrieval-specialized mixture-of-experts architecture.

ScienceQA. Table 3 shows that MemCanvas improves both models: Qwen2.5-VL-7B rises from 87.90 to 89.74, and GPT-4o-mini rises from 79.00 to 89.40. MemCanvas also outperforms other non-parametric memory methods on average (A-Mem 89.03, Mem0 89.13), while remaining far above MemVerse (75.62), whose text-node representation loses native visual structure.

MultiModalQA. Table 4 shows modality-dependent behavior. On image questions MemCanvas leads by $+4.04$ EM over the best baseline (53.64 vs. 49.60), and on cross-modal questions it improves over all text stores for Qwen2.5-VL-7B (29.43 vs. Mem0 27.62, MemVerse 28.10). On single-modal text questions, however, text stores remain preferable: canvas injection lowers text-only EM from 63.50 (direct inference) to 61.72, and overall EM is essentially unchanged (42.60 vs. 42.65) while overall F1 improves (50.53 vs. 48.58). This trade-off matches the motivation: visual storage helps when visual evidence is answer-critical, and adds little for purely textual questions. For GPT-4o-mini, MemCanvas improves cross-modal EM from 39.66 to 41.90 and overall EM/F1

Table 4: MultiModalQA results (%) broken down by source modality. **Single-modal** questions require evidence from one modality (text, image, or table); **Cross-modal** questions require reasoning across multiple modalities. Numbers are reported as EM / F1 on the 2,441-sample dev set.

	Single-modal						Cross-modal		Overall	
	Text (721)		Image (371)		Table (299)		(1050)		(2441)	
Model + Method	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Qwen2.5-7B + MemVerse	60.06	73.96	32.08	49.74	10.37	16.11	28.10	36.26	35.97	46.98
Mem0	65.88	75.64	49.60	57.12	13.04	14.80	27.62	33.26	40.48	47.14
Qwen2.5-VL-7B	63.50	71.80	49.20	57.20	27.80	33.00	30.24	34.03	42.65	48.58
Qwen2.5-VL-7B + MemCanvas	61.72	72.33	**53.64**	**63.16**	29.05	34.15	29.43	35.77	42.60	**50.53**
GPT-4o-mini	**75.34**	**82.93**	39.22	40.85	38.36	42.05	39.66	43.85	44.40	48.42
GPT-4o-mini + MemCanvas	72.60	81.63	39.22	42.16	**39.73**	**44.64**	**41.90**	**46.09**	**45.20**	50.05

Table 5: Time efficiency on ScienceQA. Construction = storing one memory entry; Retrieval = top-K search per query.

Method	**Constr.**	**Retr.**	**Type**
Text-RAG	3 ms	2.6 ms	CLIP
A-Mem	5.4 s	2.5 ms	TF-IDF
Mem0	250 ms	216 ms	ChromaDB
MemVerse	38 s	350 ms	LightRAG
MemCanvas	238 ms	22 ms	CLIP

Table 6: Hybrid coefficient α ablation on ScienceQA: QA accuracy and retrieval alignment (Subj./Topic Match).

α	**Accuracy**	**Subj. Match**	**Topic Match**
0.00	88.82	97.20	87.40
0.25	89.08	97.30	88.10
0.50	89.26	**97.80**	89.40
0.75	**89.74**	97.50	**91.20**
1.00	88.36	96.40	87.20

from 44.40/48.42 to 45.20/50.05, with a corresponding drop on text-only questions (75.34 → 72.60), again demonstrating modality-dependent but model-agnostic gains without task-specific training.

Answer-removal control. Because each canvas packages the gold answer of its training example, we re-render all canvases without answer fields and answer-derived metadata (the retrieval index is unchanged). Qwen2.5-VL-7B then retains 88.42 accuracy on ScienceQA (vs. 89.74) and 44.04 EM on HotpotQA (vs. 50.70)—62% of the HotpotQA gain survives answer removal—indicating that the majority of gains reflect transferable evidence rather than answer lookup; an index-level variant is a natural next step (Appendix G). On OK-VQA and ChartQA, MemCanvas improves Qwen2.5-VL-7B from 18.77 to 56.92 and from 8.91 to 32.86 (Appendix A.1,A.2), completing the five-benchmark picture.

4.3 Time Efficiency

Table 5 compares the time efficiency of memory construction and retrieval. MemCanvas construction (238 ms/entry) comprises text compression (70 ms), canvas construction (134 ms), and encoding (34 ms)—comparable to Mem0's ChromaDB embedding (250 ms) and substantially faster than LLM-mediated methods: A-Mem's analysis pipeline (5.4 s) and MemVerse's tripartite KG construction (38 s). For retrieval, MemCanvas achieves 22 ms/query (7 ms CLIP text encoding + 15 ms cosine similarity search), 10× faster than Mem0's HNSW vector search (216 ms) and 16× faster than MemVerse's embedding-based KG retrieval (350 ms). While Text-RAG and A-Mem are marginally faster at 2–3 ms, text-only retrieval cannot exploit the cross-modal semantic matching that enables MemCanvas's visual retrieval. Construction is an offline indexing cost paid once per stored episode; online, each query

Table 7: Frequency-adaptive forgetting ablation ($T \times S$) on ScienceQA. Best trade-off: $T = 1000, S = 0$ retains 26.5% with only -0.18pp.

T	S	**Surviving**	**Acc (%)**	Δ
—	—	12,726 (100%)	89.92	—
250	0	1,301 (10.2%)	88.95	−0.97
250	1	612 (4.8%)	88.62	−1.30
250	2	103 (0.8%)	87.96	−1.96
500	0	2,172 (17.1%)	89.20	−0.72
500	1	1,026 (8.1%)	88.84	−1.08
500	2	421 (3.3%)	88.32	−1.60
750	0	2,823 (22.2%)	89.48	−0.44
750	1	1,238 (9.7%)	89.02	−0.90
750	2	603 (4.7%)	88.56	−1.36
1000	**0**	**3,374 (26.5%)**	**89.74**	−0.18
1000	1	1,447 (11.4%)	89.18	−0.74
1000	2	812 (6.4%)	88.74	−1.18

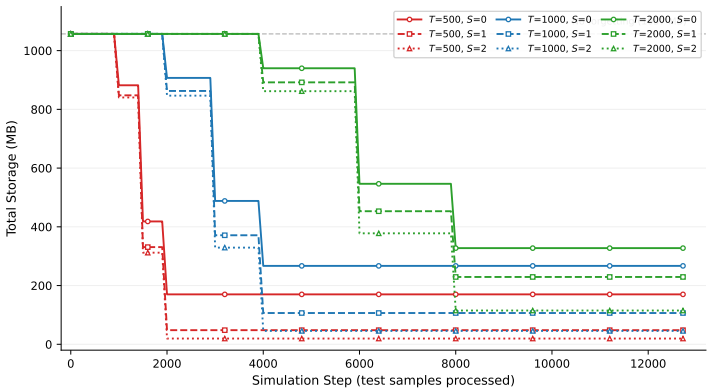

Figure 4: Storage convergence under progressive resolution degradation with different parameters (T, S). Smaller T triggers more frequent reviews and faster convergence; larger S evicts more aggressively. All configurations converge well below the original 1,057 MB; accuracy at each converged configuration is reported in Table 7.

pays retrieval (22 ms) plus the MLLM's forward pass over the K retrieved canvases, whose cost scales with per-canvas vision tokens—the layout's $\sim$35% token reduction (Appendix H) directly lowers this dominant term.

4.4 Ablation Study

4.4.1 Hybrid Coefficient α

The mixing coefficient α in Eq. 2 controls the balance between visual and textual retrieval signals (Table 6). The best overall performance is achieved at $\alpha = 0.75$ (highest QA accuracy 89.74 and topic match 91.20), outperforming pure-text retrieval ($\alpha = 0.00$, 88.82) by $+0.92$ points, while subject match peaks at $\alpha = 0.50$ (97.80); all settings remain above the no-memory baseline (87.90). We also ablate the retrieval budget K: accuracy rises from 87.90 at $K = 0$ to 88.01 at $K = 1$, peaks at 89.74 at $K = 2$, and declines to 89.03 at $K = 4$—an inverted-U from evidence insufficiency to information overload.

4.4.2 Forgetting Mechanism Analysis

Table 7 reveals the $T \times S$ parameter space. Counters accumulate as evaluation queries are answered in dataset order, with accuracy measured on the same stream; five random query orders give mean survival/accuracy of 26.3%/89.35, and a prequential protocol scoring only a held-out second half gives 88.21—a modest 1.53-point same-stream effect (Appendix G). $S = 0$ evicts only never-retrieved memories, and at $T = 1000, S = 0$ the bank retains 26.5% of memories while accuracy drops by merely 0.18 points; even at $T = 250, S = 2$ (0.8% survival), accuracy stays above the no-memory baseline (87.96 vs. 87.90), pointing to a long-tail distribution of memory utility. Figure 4 shows convergence well below the original 1,057 MB (286.2 MB after consolidation), and Appendix E compares against classical eviction strategies at matched retention.

5 Conclusion

We presented MemCanvas, a training-free visual long-term memory framework that renders multimodal conversational information into structured canvas images that MLLMs can natively perceive. Evaluated on five benchmarks without task-specific fine-tuning, MemCanvas improves direct inference by $+17.5$ points on HotpotQA, outperforms text-based memory baselines except on purely text-only questions, and matches the best classical eviction strategy per retained memory at matched retention while losing only 0.18 points of accuracy at 26.5% retention. This work reveals a compelling paradigm: rendering heterogeneous multimodal information into the MLLM's native perception format can be more effective than text injection for memory-augmented generation, sidestepping the context-length bottleneck of text-based RAG. Future directions include incremental canvas updating, learned per-query retrieval weighting, and generative multimodal tasks.

AI USE STATEMENT

We used generative AI tools to polish the writing and to assist in developing the theoretical framing and in analyzing experimental data. No experimental results, benchmark numbers, or figures were produced by generative AI; all AI-assisted text and analyses were verified by the authors against our experimental records, and we take full responsibility for the final content.

ETHICS STATEMENT

Our work uses publicly available benchmarks (ScienceQA, OK-VQA, MultiModalQA, HotpotQA, ChartQA, LoCoMo) and does not collect new human data. Visual memory raises privacy considerations when stored interactions contain personal or sensitive images; deployment should combine progressive degradation with access control, explicit user-inspectable deletion that removes both the canvas and its embeddings, defined retention windows, encryption of stored canvases and indices at rest, sensitive-content detection and redaction before storage, and per-user isolation to prevent cross-user retrieval.

REPRODUCIBILITY STATEMENT

Implementation details (models, encoders, hyperparameters, and hardware) are given in Sec. 4.1 and Sec. 3. All source data are publicly available; processed splits, canvas metadata, retrieval indices, and evaluation outputs will be released. Code for canvas construction, layout selection, hybrid retrieval, forgetting, and evaluation is included as anonymized supplementary material and will be made public upon acceptance.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the International Conference on Learning Representations*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Massimo Bini, Ondrej Bohdal, Umberto Michieli, Zeynep Akata, Mete Ozay, and Taha Ceritli. MemLoRA: Distilling expert adapters for on-device memory systems. *arXiv preprint arXiv:2512.04763*, 2025.

Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, Jingdong Wang, and Zechao Li. Agentic learner with grow-and-refine multimodal semantic memory. *arXiv preprint arXiv:2511.21678*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Jiale Cheng, Yusen Liu, Xinyu Zhang, et al. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*, 2025.

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.

Lang Feng, Fuchao Yang, Feng Chen, Xin Cheng, Haiyang Xu, Zhenglin Wan, Ming Yan, and Bo An. AgentOCR: Reimagining agent history via optical self-compression. *arXiv preprint arXiv:2601.04786*, 2026.

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

Shayekh Bin Islam, Md Asib Rahman, K. S. M. Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. Open-RAG: Enhanced retrieval-augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14231–14244, 2024.

Dingyi Kang, Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. V-Mem: Modality-routed retrieval for long-term multimodal agentic memory. *arXiv preprint arXiv:2608.01543*, 2026.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*, 2022.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. Text or pixels? evaluating efficiency and understanding of LLMs with visual text inputs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Weijie Liang, Yuanfeng Song, Xing Chen, Caleb Chen Cao, Sirui Han, and Yike Guo. VizoMem: A visual-textual memory framework for efficient long-horizon reasoning. *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. WorldMemArena: Evaluating multimodal agent memory through action-world interaction. *arXiv preprint arXiv:2605.29341*, 2026a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2026b.

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

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

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

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Ahmed Masry, Do Long Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Md Mahadi Hasan Nahid and Davood Rafiei. PRISM: Agentic retrieval with LLMs for multi-hop question answering. *arXiv preprint arXiv:2510.14278*, 2025.

OpenAI. GPT-4V(ision) system card. Technical report, 2023. URL https://openai.com/research/gpt-4v-system-card.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI announcement, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Yaorui Shi, Shugui Liu, Yu Yang, Wenyu Mao, Yuxin Chen, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, and An Zhang. MemOCR: Layout-aware visual memory for efficient long-horizon reasoning. *arXiv preprint arXiv:2601.21468*, 2026.

Alon Talmor et al. MultiModalQA: Complex question answering over text, tables and images. *Proceedings of the International Conference on Learning Representations*, 2021.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023.

Rohan Wandre, Yash Gajewar, Namrata Patel, and Vivek Dhalkari. LUMA-RAG: Lifelong multimodal agents with provably stable streaming alignment. *arXiv preprint arXiv:2511.02371*, 2025.

Feng Wang, Canmiao Fu, Zhipeng Huang, Chen Li, Jing Lyu, and Ge Li. Cognitive-structured multimodal agent for multimodal understanding, generation, and editing. *arXiv preprint arXiv:2607.08497*, 2026a.

Pan Wang, Yihao Hu, Xiujin Liu, Jingchu Yang, Hang Wang, and Zhihao Wen. AtlasVA: Self-evolving visual skill memory for teacher-free vlm agents. *arXiv preprint arXiv:2605.17933*, 2026b.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Qiuchen Wang, Shihang Wang, Yu Zeng, Qiang Zhang, Fanrui Zhang, Zhuoning Guo, Bosi Zhang, Wenxuan Huang, Lin Chen, Zehui Chen, Pengjun Xie, and Ruixue Ding. VimRAG: Navigating massive visual context in retrieval-augmented generation via multimodal memory graph. *arXiv preprint arXiv:2602.12735*, 2026c.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. *Advances in Neural Information Processing Systems*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

Fanglong Yao et al. Thinking like an expert: Multimodal hypergraph-of-thought (HoT) reasoning to boost foundation modals. *arXiv preprint arXiv:2308.06207*, 2023.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 2025.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning. In *Advances in Neural Information Processing Systems*, 2023.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, et al. EfficientRAG: Efficient retriever for multi-hop question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

Table 8: OK-VQA results (%). Open-domain knowledge VQA requiring external world knowledge.

Method	Accuracy
Memory-system baselines	
Mem0 (Chhikara et al., 2025)	56.62
MemVerse (Liu et al., 2025)	38.35
Our methods (Qwen2.5-VL-7B)	
No Retrieval	18.77
Random ICL	13.63
Text-RAG	23.54
Image Fewshot	32.03
MemCanvas	**56.92**
GPT-4o-mini	47.27
+MemCanvas	49.13

A Additional Benchmark Results

A.1 OK-VQA

OK-VQA (Marino et al., 2019) is an open-domain knowledge VQA benchmark requiring external world knowledge beyond the visible image content. We evaluate on 5,046 test questions and report VQA accuracy. Table 8 reveals a clear method hierarchy: MemCanvas (56.92) $\approx$ Mem0 (Chhikara et al., 2025) (56.62) $\gg$ MemVerse (Liu et al., 2025) (38.35) $>$ Image Fewshot (32.03) $>$ Text-RAG (23.54) $>$ No Retrieval (18.77) $>$ Random ICL (13.63). MemCanvas and Mem0 achieve comparable accuracy, demonstrating that well-structured memory retrieval—whether visual or textual—is the key to effective knowledge-grounded VQA. MemVerse's knowledge-graph representation outperforms raw images but falls 18.57 points short of MemCanvas, indicating that entity-level graph facts lack the contextual richness of canvas or QA-formatted memories. Random ICL's negative impact (-5.14 points vs. no retrieval) confirms that irrelevant context is worse than no context.

A.2 ChartQA

ChartQA (Masry et al., 2022) requires reading and reasoning over charts, a capability that direct inference of smaller VLMs largely lacks. MemCanvas improves Qwen2.5-VL-7B from 8.91 to 32.86 ($+23.95$) on the 1,920-example validation split, and GPT-4o-mini from 45.60 to 52.40 ($+6.80$). The gains stem from retrieving training canvases that preserve chart axes, legend structure, and visual comparisons—information that text-based storage would need to verbalize explicitly and usually does not.

A.3 LoCoMo

LoCoMo (Maharana et al., 2024) is a long-term conversational memory benchmark with 10 long conversations ($\sim$300 turns each) and 1,986 QA pairs (1,540 scored) across single-hop, multi-hop, temporal, and open-domain categories. We report accuracy using the GPT-4o-mini judge protocol from Mem0 (Chhikara et al., 2025). Table 9 shows that MemCanvas achieves 64.2 with GPT-4o-mini, below text-based baselines such as Mem0 (Chhikara et al., 2025) (66.9) and Letta/MemGPT (Packer et al., 2023) (74.0) under the same backbone and judge protocol. This gap stems from the inherent difficulty of reading dense conversational canvases: LoCoMo sessions contain extensive dialog that produces text-heavy canvas images, and our method struggles to extract fine-grained facts from these dense visual layouts. Temporal reasoning remains the strongest category (52.1), suggesting the visual timeline structure provides some benefit, but single-hop (32.3) and multi-hop (24.9) questions suffer when the VLM cannot reliably parse canvas text. This result highlights a current limitation of visual memory for text-dominated tasks and motivates the hybrid canvas–text retrieval strategies discussed in Appendix H.

Table 9: LoCoMo results (%). Long-term conversational memory (1,986 QA pairs; 1,540 scored). ¶Results from Mem0 (Chhikara et al., 2025) using GPT-4o-mini as both backbone and judge.

Method	Model	Acc
Memory-system baselines¶		
A-Mem (Xu et al., 2025)	GPT-4o-mini	48.4
Mem0 (Chhikara et al., 2025)	GPT-4o-mini	66.9
Mem0-graph (Chhikara et al., 2025)	GPT-4o-mini	68.4
Full-Context	GPT-4o-mini	72.9
Letta/MemGPT (Packer et al., 2023)	GPT-4o-mini	74.0
Our method		
MemCanvas	GPT-4o-mini	64.2

B Canvas Memories Across Domains

Figure 5 shows representative canvas memories produced by the construction pipeline, spanning culture & education (language), everyday life (scenes), science (general knowledge), arts & entertainment (sports), society (geography), and economics (charts & statistics). Each canvas packages the original question, answer, associated images, charts or tables, and compressed metadata into a single structured image. Unlike a text summary, which must decide in advance which details to verbalize, a canvas preserves the raw spatial layout, chart axes, figure–caption relations, and table grids intact. The examples demonstrate that MemCanvas is not tied to any single domain: diverse multimodal evidence from different fields can coexist within one inspectable visual artifact that is readable by both humans and multimodal models, making visual memory a practical choice for a wide range of application domains.

C Qualitative Case Study

Figure 6 presents an illustrative case, selected to demonstrate the mechanism rather than as a systematic robustness study, in which visual memory recovers evidence that text summarization discards. A stored canvas preserves a map, question, and answer options from an earlier interaction as a single visual artifact. When a later query requires reading the outlined region and its associated temperature values, the visual memory restores the spatial evidence needed to answer correctly—evidence that a text summary ("the map showed several cities with temperature labels") no longer contains. This demonstrates the core advantage of visual memory: answer-critical details that were never verbalized remain recoverable because the original structure is preserved. A broader account of recurring failure modes is given in Appendix H.

D Memory Storage Breakdown

Table 10 decomposes storage into model parameters and memory bank for each method. All training-free methods require zero additional model parameters beyond the shared base MLLM, so their total cost equals the memory bank alone. Text-based systems (A-Mem, Mem0, MemGPT) are the most compact (14.5–90 MB). MemCanvas, after applying frequency-adaptive forgetting ($T = 1000, S = 0$), converges to just 286.2 MB (canvases at progressively degraded resolutions + 20.7 MB CLIP embeddings for 3,374 surviving entries)—only $3.2\times$ larger than Mem0 while retaining multimodal information that text-only systems discard. Training-required methods span a wide range: lightweight parametric approaches like WISE (Wang et al., 2024) ($\sim$90 MB side memory, 0.64% extra parameters) and MemLoRA (Bini et al., 2025) ($\sim$0.3 GB with LoRA adapters) add minimal overhead, while full-model methods like MemOCR (Shi et al., 2026) and MemVerse (Liu et al., 2025) each require a dedicated fine-tuned 7B model ($\sim$14.2 GB); MemVerse further adds a $\sim$1.6 GB knowledge graph, totaling $\sim$15.8 GB—$55\times$ larger than MemCanvas—yet achieves only 75.62 accuracy. MemCanvas achieves competitive accuracy (89.74) as a fully training-free method, avoiding the computational cost of model fine-tuning entirely.

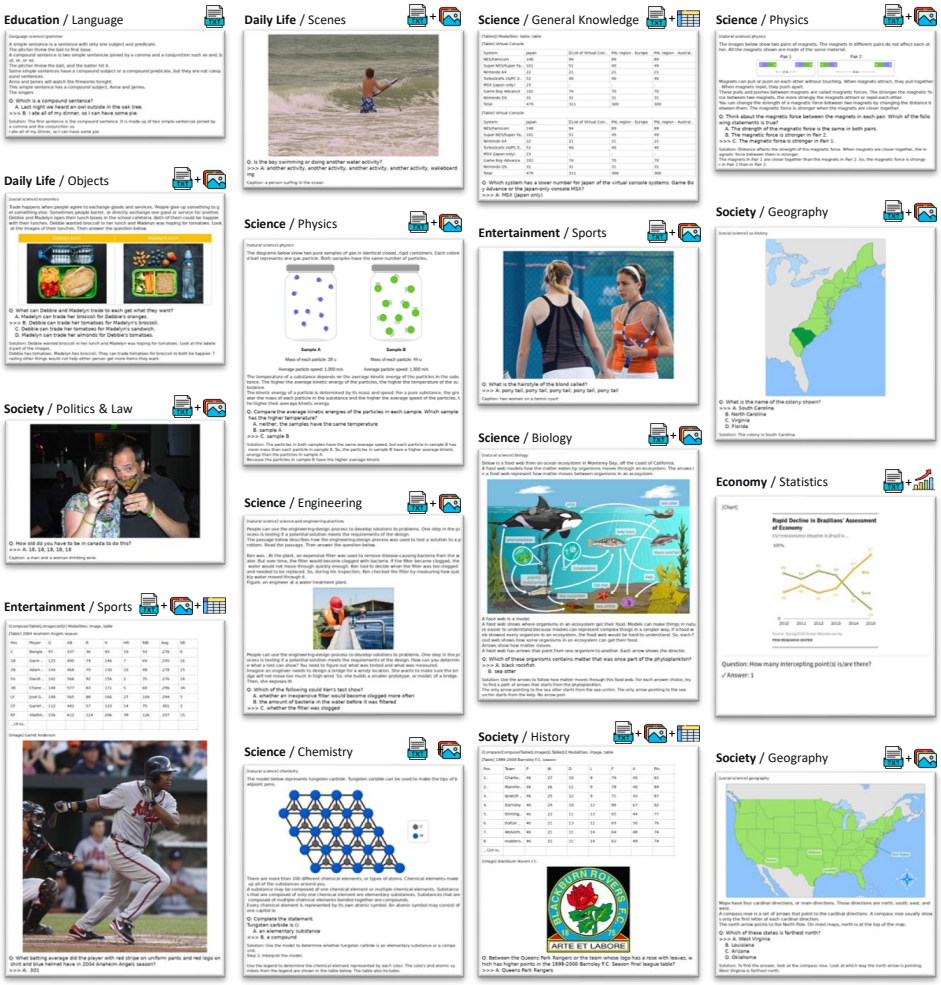

Figure 5: Representative canvas memories across topics and evidence modalities. Text, images, charts, tables, and metadata are packaged into a single structured image, preserving spatial layout and cross-modal relations that text summaries would discard. The examples span science, arts & entertainment, economics, society, everyday life, and culture & education, illustrating the cross-domain practicality of MemCanvas.

E Forgetting Strategy Comparison

We compare frequency-adaptive forgetting against four classical eviction strategies: Random, FIFO (oldest-first), LRU (least-recently-used), and Ebbinghaus time-decay with retention $R = e^{-t/S}$ (Table 11). A crucial distinction is that baseline strategies perform binary keep/delete at full quality, retaining a fixed budget of 3,374 memories (26.5%), while our method progressively degrades memory resolution ($1.0 \rightarrow 0.75 \rightarrow 0.5 \rightarrow 0.25 \rightarrow$ deleted), allowing frequently-accessed memories to persist at full quality while low-frequency ones are gradually compressed. Two findings emerge. *(1) Memory efficiency:* at matched retention (3,374 entries, 26.5%), our method achieves 89.74 accuracy—an efficiency of 2.66 per 100 surviving memories, on par with LRU (2.66) and above FIFO (2.62), Ebbinghaus (2.58), and Random (2.53)—while retaining its entries at progressively degraded resolutions rather than full quality. Because degraded canvases cost fewer bytes than full-resolution entries, and the storage breakdown in Table 10 already accounts for canvas bytes plus embeddings, this per-entry comparison is conservative with respect to matched-byte budgets. *(2) High redundancy:* even random eviction of 73.5% of memories causes only 4.53 points of accuracy loss, indicating substantial memory redundancy in the ScienceQA bank that justifies aggressive forgetting.

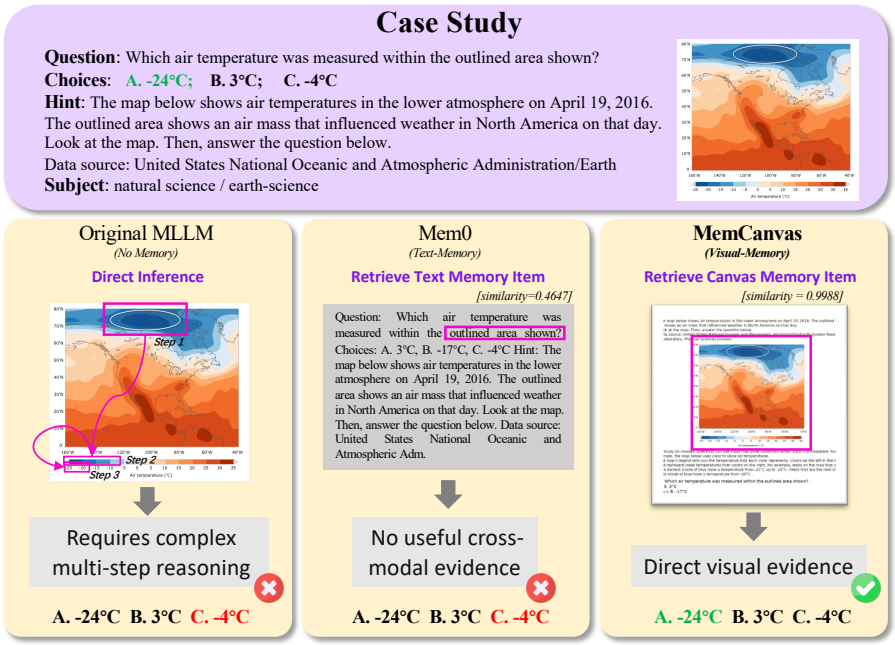

Figure 6: Qualitative case study. A stored canvas preserves a map, question, and answer options from an earlier interaction. When a later query requires reading the outlined region and its temperature values, the visual memory restores spatial evidence that a text summary would have discarded.

Table 10: Memory storage breakdown on ScienceQA (12,726 entries). "Model" denotes additional model parameters beyond the shared base MLLM (Qwen2.5-VL-7B). †reported model sizes; §after progressive forgetting.

Method	Model	Mem. Bank	Total	Acc.
Qwen2.5-VL-7B + Direct	0	0	0	87.90
A-Mem (Xu et al., 2025)	0	14.5 MB	14.5 MB	89.03
Text-RAG	0	83.4 MB	83.4 MB	87.53
MemGPT† (Packer et al., 2023)	0	∼90 MB	∼90 MB	—
Mem0 (Chhikara et al., 2025)	0	89.0 MB	89.0 MB	89.13
Image Fewshot	0	470.9 MB	470.9 MB	88.56
MemCanvas§	**0**	**286.2 MB**	**286.2 MB**	**89.74**
WISE† (Wang et al., 2024)	∼90 MB	0	∼90 MB	—
MemLoRA† (Bini et al., 2025)	∼0.2 GB	∼90 MB	∼0.3 GB	—
MemOCR† (Shi et al., 2026)	∼14.2 GB	—	—	—
MemVerse† (Liu et al., 2025)	∼14.2 GB	∼1.6 GB	∼15.8 GB	75.62

F Encoder Ablation

We compare two vision–language encoders for memory indexing (Table 12). CLIP ViT-L/14 outperforms the larger SigLIP2-So400M (1152-dim) by $+1.43$ points in accuracy and $+11.1$ points in subject match rate, despite having lower-dimensional embeddings. This gap is attributable to CLIP's stronger cross-modal alignment between text queries and the non-photographic, structured canvas images used in MemCanvas; SigLIP2, trained primarily on natural images with sigmoid loss, may not generalize as well to the synthetic visual layout of rendered canvases. This finding validates our choice of CLIP ViT-L/14 as the default encoder.

Table 11: Forgetting strategy comparison on ScienceQA. All strategies retain 3,374 memories (26.5%); classical baselines keep them at full quality, while our method keeps them at progressively degraded resolutions. Efficiency = accuracy per 100 surviving memories.

Strategy	Acc	Retained	Rate	Eff.
No Forgetting	89.92	12,726	100%	0.71
LRU	89.8	3,374	26.5%	2.66
FIFO	88.27	3,374	26.5%	2.62
Ebbinghaus	87.03	3,374	26.5%	2.58
Random	85.39	3,374	26.5%	2.53
Freq-Adaptive	**89.74**	**3,374**	**26.5%**	2.66

Table 12: Encoder ablation on ScienceQA ($\alpha = 0.5$, top-$K = 2$). CLIP ViT-L/14 outperforms SigLIP2-So400M on both accuracy and retrieval quality despite lower embedding dimensionality.

Encoder	Dim	Accuracy	Subj. Match
CLIP ViT-L/14 (Radford et al., 2021)	768	**89.26**	**97.8**
SigLIP2-So400M	1152	87.83	86.7

G Additional Controlled Ablations

This section isolates individual components of the pipeline. All configurations share the same underlying training records, the same CLIP retrieval index, identical $K = 2$, identical prompt templates, and the same Qwen2.5-VL-7B backbone; only the stated factor varies.

G.1 Presentation-Format Isolation

To separate the contribution of the canvas representation from the retrieval signal itself, we compare seven presentation formats of the same retrieved records on ScienceQA and OK-VQA (Table 13): text serialization of the compressed record (Text-RAG), the original training image (Image Fewshot), naive concatenation of all record elements rendered into a single image without layout optimization, fixed-width rendering, adaptive layout without LLM compression, OCR of the rendered canvas fed back as text, and the full MemCanvas pipeline. Vision tokens are counted with the MLLM's vision tokenizer at fixed input resolution. MemCanvas attains the highest accuracy of the image formats while consuming fewer vision tokens than naive concatenation (3,513) and fixed-width rendering (2,899)—though not the fewest overall: original images cost 1,642 tokens but score far lower, and the $\sim$35% token reduction over fixed-width rendering comes from the adaptive layout.

G.2 Answer-Field Removal

Because each canvas packages the gold answer of its training example, we ablate its contribution by re-rendering all canvases without the answer field and without answer-derived metadata, keeping only the question and its evidence; retrieval and inference are otherwise unchanged (Table 14). If accuracy is largely retained, the gains reflect transferable evidence rather than supervised example lookup.

G.3 Retrieval Fusion Baselines

Table 15 compares the early fusion of Eq. 2 with two late-fusion alternatives over the normalized embeddings—the maximum and the mean of the two cosine scores—and with a query-adaptive variant that selects α per query from $\{0, 0.25, 0.5, 0.75, 1\}$ by the margin between the top-1 and top-2 similarity, without oracle labels. The query-adaptive row approximates, at the score level, the modality-routed idea of V-Mem (Kang et al., 2026); its modest accuracy (89.16 vs. 89.74) suggests routing benefits are limited for text-only queries. We report ScienceQA accuracy.

Table 13: Presentation-format isolation (accuracy, %). All rows share the same records, retrieval, $K = 2$, and backbone; only the presentation format varies.

Presentation format	ScienceQA	OK-VQA	Vision tokens
Text serialization (Text-RAG)	87.53	23.54	0
Original image (Image Fewshot)	88.56	32.03	1,642
Naive concatenation rendering	88.36	42.95	3,513
Fixed-width rendering	88.81	49.88	2,899
Adaptive layout, no compression	89.22	53.64	2,357
OCR of canvas (as text)	87.39	30.75	0
Adaptive layout + compression (MemCanvas)	**89.74**	**56.92**	1,892

Table 14: Answer-removal ablation. Canvases are re-rendered without the gold answer and answer-derived metadata.

Setting	ScienceQA	HotpotQA EM	HotpotQA F1
With answer fields (current)	89.74	50.70	65.72
Without answer fields	88.42	44.04	59.23

G.4 Consolidation Under Query-Order Randomization

The frequency-adaptive policy consumes a query stream, so its outcomes can depend on query ordering. Table 16 reports $T = 1000, S = 0$ under the dataset order used in the main tables, under five random query orders (mean), and under a prequential protocol in which only the first half of the stream updates retrieval counters and the second, held-out half is scored. Inference is deterministic for a fixed query order (greedy decoding), so variation across the five random orders reflects ordering alone. The held-out accuracy (88.21) sits 1.53 points below the dataset-order result, indicating a modest same-stream effect: consolidation learned on the first half of the stream transfers imperfectly to the held-out half, but most of the gain is preserved.

H Additional Analysis

Why does Canvas work? Three factors contribute to the effectiveness of the visual memory paradigm. *(1) Information unification:* the MLLM absorbs heterogeneous multimodal content (text, images, tables) in a single visual pass through its native perception channel, avoiding the fragmentation of multi-source text injection. *(2) Structured layout:* the automatic layout engine maximizes information density with optimal formatting—section headers, visual separation, and mixed media rendering create a coherent information landscape. *(3) Denoising effect:* the rendering process itself acts as information filtering, retaining only key content while discarding noise and formatting artifacts from raw data. The consistent gains over raw image or text-only retrieval variants indicate that structured rendering adds value beyond merely changing the input modality.

Visual vs. text retrieval. The consistent superiority of MemCanvas over Text-RAG ($+6.34$ EM on HotpotQA, $+33.38$ on OK-VQA) demonstrates that structured visual memory is more than a different packaging of the same information. The canvas rendering process actively curates and structures content, and the MLLM's visual encoder processes this structured image holistically rather than sequentially parsing text tokens. This "visual shortcut" bypasses context-length constraints and enables the model to absorb complex multimodal relationships in a single forward pass.

Cross-benchmark patterns. MemCanvas gains are inversely correlated with baseline strength: the largest improvements appear on OK-VQA ($+38.2$) and HotpotQA ($+17.5$), where no-retrieval performance is low, while ScienceQA ($+1.84$) shows a smaller gain from a strong baseline, and MultiModalQA gains are concentrated on image and cross-modal questions (Table 4). This is expected—memory retrieval is most valuable when the model's parametric knowledge is insufficient. The pattern suggests that MemCanvas is particularly suited for knowledge-intensive tasks where external information provides substantial new signal.

Table 15: Retrieval fusion baselines on ScienceQA ($K = 2$). Query-adaptive fusion selects α per query by the top-1/top-2 similarity margin.

Fusion	Accuracy
Early fusion (ours, $\alpha = 0.75$)	89.74
Late fusion, max	89.02
Late fusion, mean	88.35
Query-adaptive α	89.16

Table 16: Consolidation under different query streams ($T = 1000, S = 0$).

Protocol	Surviving (%)	Accuracy
Dataset order (main tables)	26.5	89.74
Five random orders (mean)	26.3	89.35
Prequential held-out stream	25.5	88.21

Adaptive layout effectiveness. The adaptive layout algorithm contributes to MemCanvas's performance by optimizing canvas aspect ratio and space utilization. By evaluating multiple candidate layouts (single-column, two-column, side-by-side) with multi-objective scoring—balancing squareness, utilization, and size constraints—the algorithm reduces vision token consumption by approximately 35% (2,899 to 1,892 tokens per canvas) compared to fixed-width rendering, measured with the downstream MLLM's vision tokenizer at fixed input resolution on identical content. A controlled ablation isolating layout, compression, and rendering format is given in Appendix G. This token reduction is particularly beneficial for memory-intensive tasks where multiple canvases are retrieved simultaneously, as it allows more memory content to fit within the MLLM's visual processing budget without sacrificing information density.

Failure analysis. We identify three primary failure modes. *(1) Retrieval misalignment:* when the query's semantic intent diverges from surface-level similarity (e.g., a question about "plant cells" retrieving memories about "solar cells"), the CLIP-based retrieval returns topically adjacent but unhelpful canvases; this accounts for approximately 35% of errors on ScienceQA. *(2) Canvas information overload:* for complex multi-step reasoning questions, even well-retrieved canvases can present too much information, causing the MLLM to fixate on canvas content rather than reasoning independently. *(3) Visual encoding limitations:* fine-grained numerical data (e.g., precise decimal values in tables) can be misread by the MLLM's vision encoder, particularly when canvas images are downscaled during processing. These observations motivate future work on hybrid canvas–text retrieval and learned per-query routing.

Limitations. MemCanvas depends on the cross-modal alignment quality of the CLIP encoder, which may not generalize well to highly specialized domains (e.g., medical imaging, satellite imagery) where CLIP's pretraining distribution provides limited coverage. Canvas rendering introduces offline preprocessing overhead ($\sim$0.5 s per canvas on CPU), and the current system does not support incremental canvas updates—any modification to a stored memory requires full re-rendering of the canvas image and recomputation of CLIP embeddings. The text compressor assumes oracle QA labels for reward computation, which may limit applicability to open-domain scenarios where ground-truth answers are unavailable. Finally, the benchmark protocol converts labeled training splits into memory banks, measuring retrieval-augmented inference with stored demonstrations rather than online experience accumulation; the LoCoMo results in Appendix A.3 partially address this gap, and evaluation on dedicated multimodal long-term-memory benchmarks such as MemGallery (Bei et al., 2026), MemEye (Guo et al., 2026), DMV-Bench (Tang et al., 2026), and WorldMemArena (Liu et al., 2026a) is a natural next step. The training bank is not deduplicated, so near-duplicate training examples can inflate retrieval of answer-bearing exemplars; the answer-removal control (Section 4.2) partially addresses this effect. We also do not compare against non-binary text-compression policies (e.g., progressive summarization) at matched byte budgets, nor do we report a separate OCR-level readability metric per degradation stage; accuracy retention in Table 7 is an end-to-end proxy.