INTUITION WITH EVIDENCE: ADAPTIVE DUAL MEMORY FOR MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

As multimodal agents engage in sustained interactions, long-term memory serves two distinct but complementary functions: fine-grained evidence recall and internalized interpretation for global understanding. Although episodic and latent memory address these demands through explicit historical records and internalized experience, neither paradigm alone can accommodate the diverse memory affinities across tasks, owing to limitations in episodic evidence coverage and integration, and in latent factual grounding. To bridge this divide, we propose a dual-memory framework, unifying a model-native latent memory as an intuitive prior with a multi-indexed episodic memory for exact evidence reconstruction via retrieval-conditioned joint alignment. At inference, the framework adaptively composes these memories through a training-free routing mechanism driven by Jensen-Shannon Divergence (JSD) sensitivity and confidence proxies for query-dependent memory affinity. Extensive evaluations across multimodal benchmarks demonstrate state-of-the-art performance, outperforming single-paradigm baselines overall (+4.67% over vanilla Episodic), with further analysis highlighting episodic strengths in evidence-localized tasks and latent contributions to integrating dispersed information and compensating for imperfect retrieval.

1 INTRODUCTION

As multimodal agents transition to sustained interactions, their long-term memory requirements diversify to encompass highly distinct but complementary objectives including precise recall of verifiable historical evidence (Wu et al., 2024) and internalization of cross-session experiences into intuitive understanding (Park et al., 2023). Although effective reasoning relies on both, they impose fundamentally divergent organizational and computational paradigms: exact recall demands recoverable and discrete records to ensure factual fidelity, while contextual reasoning requires abstracting sparse trajectories into a compressed understanding, seamlessly aligned with the model's native comprehension capabilities (Joren et al., 2025).

Existing single-paradigm architectures fail to satisfy both demands. While episodic memory offers interpretable retrieval and extracts semantics via hierarchical reflection (Packer et al., 2023; Jiang et al., 2026), explicit retrieval risks incomplete evidence (Wu et al., 2024) and context dilution (Liu et al., 2024). Critically, even given sufficient information, models struggle to integrate disjointed records for organic reasoning (Amiraz et al., 2025). Conversely, while latent memory internalizes histories into continuous representations for efficient cross-instance transfer (Feng et al., 2026b; Zhang et al., 2026a), this compression inherently suffers from information entangles (Zhang et al., 2026b) without targeted record patching (Wang et al., 2024).

Crucially, real-world interactions exhibit highly diverse *memory affinities*, varying in their demand for specific facts versus global understanding. Thus, no single paradigm suffices for long-horizon tasks. Recent work coordinates episodic and latent memories mainly by converting explicit records into parametric weights via exclusive routing (Liu et al., 2025b; Xia et al., 2026), ignoring the most need for concurrent, synergistic collaboration. These limitations raise: *Can diverse mechanisms be unified regarding characteristics for varying query-specific memory affinities?*

We propose a dual-memory framework that synergizes *cognitive intuition* with *factual verification* by decoupling underlying memory roles. Specifically, a multi-indexed episodic store retrieves multimodal observations for precise *evidence grounding*, whereas an inference-side native latent memory

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.

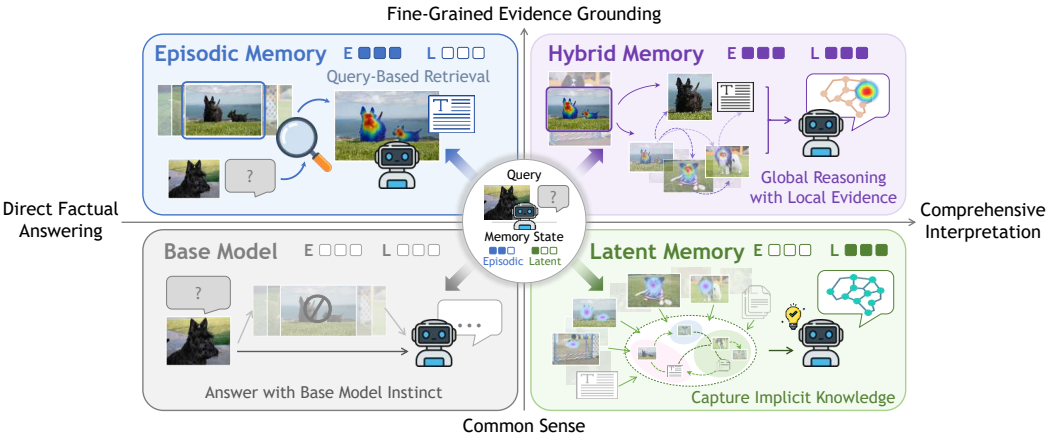


Figure 1: **Complementary memory paradigms for long-term agent reasoning.** Episodic memory reconstructs source-grounded evidence for factual recall, while Latent memory internalizes trajectories into into model-native representations for global reasoning. As neither is universally dominant, our framework maintains both mechanisms and adaptively composes them at inference.

internalizes trajectories via test-time adaptation to establish an *intuitive prior*. Rather than operating as isolated components, they are synchronized through joint alignment training for cohesive reasoning. During inference, further to dynamically match query-specific memory affinities conditioned on the current memory state, we deploy a training-free adaptive router driven by JSD sensitivity and sequence confidence proxies, unifying these mechanics into an organic memory framework.

We evaluate our framework on multi-turn multimodal benchmarks across various model scales to validate the effectiveness of hybrid memory orchestration. Beyond notable inference efficiency and robustness, our analysis uncovers distinct task-specific memory affinities, underscoring the necessity of co-utilizing complementary memory mechanisms organically. Our contributions are:

- We systematically analyze the inherent failure modes of single-paradigm memories (episodic vs. latent vs. fixed hybrid) and uncover distinct task-specific memory affinities.
- We propose a novel dual-level memory framework combining episodic evidence grounding with native latent intuitive priors, adaptively orchestrated by adaptive proxy-guided routers.
- We conduct extensive experiments to validate the effectiveness, robustness, and efficiency of the hybrid memory mechanism with in-depth analysis for memory characteristics.

2 Related Work

Episodic Memory. LLM agents commonly use external textual memory by incorporating retrieved records into the input context (Park et al., 2023). To support efficient access, these records are organized into structured stores, including hierarchies and graphs, with retrieval performed before generation or interleaved with reasoning (Chhikara et al., 2025; Xu et al., 2026; Yan et al., 2026). Multimodal extensions combine textual abstractions with visual representations, supporting retrieval of both semantic information and fine-grained perceptual evidence (Long et al., 2026; Kang et al., 2026). However, query-driven retrieval can miss implicitly relevant memories, while broader retrieval increases inference costs (Hu et al., 2026a;b). Even accurate retrieval may leave a semantic gap between human-readable representations and the model's in-context interpretation and reasoning (Wu et al., 2024; Shen et al., 2026).

Latent Memory. Latent memory encodes experience in model-native representations rather than explicit records. Some retrieve cached continuous representations or compress retrieved experience into soft tokens for inference (Feng et al., 2026b; Fu et al., 2026), while retaining dependence on retrieval coverage. Other internalize experience through training, learning either context-dependent memory generators (Zhang et al., 2026a) or reusable procedural and tool knowledge (Wu et al.,

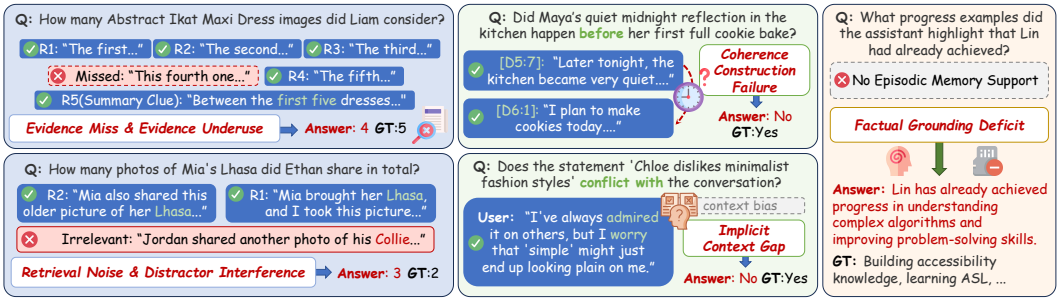


Figure 2: **Failure cases of individual memory branches conducted on Mem-Gallery (Bei et al., 2026).** *Left*: Episodic memory is limited by retrieval quality (omissions and distractions). *Middle*: Episodic memory is limited by contextual understanding (evidence integration and implicit alignment). *Right*: Latent memory fails to recover the specific details required by the query.

2026; Wang et al., 2025) without light adaptation to new scenarios. Multimodal variants generate latent perceptual and semantic memories (Yu et al., 2026), or verbalize parametric recall, retaining dependence on in-context reasoning (Liu et al., 2025a). However, preserving broad semantic content does not ensure factual fidelity as distorting details under capacity constraints (Zhang et al., 2026b).

Test-Time Adaptation. Test-time learning enables models and agents to adapt to incoming data and feedback during deployment. Early work studies self-supervised parameter updates (Sun et al., 2020), with recent extensions to LLMs (Feng et al., 2026a). It further supports context retention through recurrent and hierarchical neural memories (Behrouz et al., 2026b;a), as well as compression into adapters (Chen et al., 2026). For agents, test-time adaptation can also occur at the harness level by accumulating reflections and revising external memory without weights drifting (Shinn et al., 2023; Xu et al., 2026). These developments motivate coordinated adaptation across parametric and external memory. UniMem consolidates recurring episodic experience into parametric procedural memory, but its mutually exclusive routing ignores the characteristics between two memory (Xia et al., 2026). Building on these works, we study how combining the complementary dual-level memories can improve test-time adaptation.

3 Rethinking Long-Term Memory: Evidence or Intuition

3.1 Motivation: Uni Memory Failures in the Memory Affinity Gap

Episodic and latent memory, two major memory paradigms, operate through distinct mechanisms and have their own limitations, as illustrated by the case studies in Figure 2.

Episodic memory can fail in both evidence coverage and interpretation. On the retrieval side, beyond missing direct evidence, query similarity matching can omit supporting cross-episode background and relations (Gutiérrez et al., 2024). Broadening retrieval without adequate context may introduce distraction (Amiraz et al., 2025). On the inference side, even with sufficient evidence, the model may fail to organize and combine it correctly (Joren et al., 2025). Interpretation may further require drawing on implicit global semantics and stylistic conventions distributed across the history.

Latent memory internalizes experience within a limited parameter budget. Semantic abstraction may blur distinctions between episodes, hindering recall of the specific entities and events relevant to a query. Even when the relevant experience is recalled, the model may reinterpret and stylistically reformulate its content, compromising recovering to the original evidence. These failures suggest proper combination of distinct memory for complementary.

3.2 Method: Episodic Memory as an Evidence Layer

Episodic memory $\mathcal{E}$ supports inference with query-relevant textual and visual evidence. We organize past trajectories into episodes, each containing original records $r \in \mathcal{R}$ augmented with derived

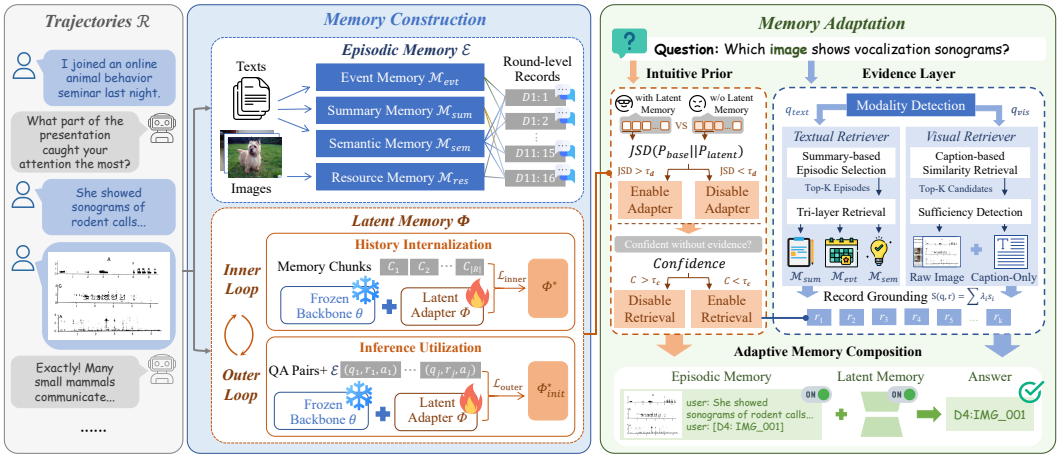


Figure 3: **Overall Workflow of our method.** Memory construction transforms multimodal trajectories into complementary episodic and latent memories. Given a query, memory adaptation selectively activates and integrates these memories by training-free routers to generate the answer.

memories $\mathcal{M}$ at multiple granularity and across modalities that serve as retrieval indices:

$$\mathcal{E} = (\mathcal{R}, \mathcal{M}), \qquad \mathcal{M} = \{\mathcal{M}_i\}_{i \in \{\text{sum,evt,sem,res}\}}. \tag{1}$$

Summary memory $\mathcal{M}_{\text{sum}}$ groups summaries, topics, and entities by chunk. Event and Semantic memories $\mathcal{M}_{\text{evt}}$ and $\mathcal{M}_{\text{sem}}$ organize event summaries and cross-episode facts and preferences, respectively, into semantic clusters. Resource memory $\mathcal{M}_{\text{res}}$ maintains a flat collection of images. Source links connect these derived memories to their supporting original records.

Given a query q containing text and an optional image, a modality router selects text-only or multimodal retrieval. Text retrieval combines matching within episodes identified via $\mathcal{M}_{\text{sum}}$ with global searches over $\mathcal{M}_{\text{sem}}$ and $\mathcal{M}_{\text{evt}}$. When required, image retrieval matches q against images in $\mathcal{M}_{\text{res}}$. Matched $\mathcal{M}$ contribute relevance scores s to the corresponding r. Each r then receives a combined score $S(q, r)$ across retrieval channels:

$$S(q,r) = \lambda_{\text{sum}} s_{\text{sum}}(q,r) + \lambda_{\text{evt}} s_{\text{evt}}(q,r) + \lambda_{\text{sem}} s_{\text{sem}}(q,r) + \lambda_{\text{res}} s_{\text{res}}(q,r), \tag{2}$$

where λ weights each channel. The top-K r ranked by $S(q, r)$ constitute the evidence set $\mathcal{R}_q$. For visual retrieval, semantic captions are selectively supplemented with original images, adaptively allocating additional budget to fine-grained visual evidence.

3.3 Method: Latent Memory as an Intuitive Prior

Reusable, high-level understanding of the past can be internalized in model parameters as latent memory, providing intuitive guidance. Latent memory is encoded in a trainable adapter ϕ attached to the frozen inference backbone θ. A meta-learning framework is adopted to enable rapid history encoding through test-time training (Sun et al., 2020; Chen et al., 2026).

First, $\mathcal{R}$ is internalized in ϕ via an inner loop that minimizes a causal language-modeling objective:

$$\mathcal{L}_{\text{inner}}(\phi, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{r=1}^{|\mathcal{R}|} \mathcal{L}_{\text{NTP}}(\phi; C_r), \tag{3}$$

where $\{C_r\}_{r=1}^{|\mathcal{R}|}$ are chunks of $\mathcal{R}$, with images represented by captions as semantic records, and $\mathcal{L}_{\text{NTP}}$ denotes the next-token prediction loss. After the inner-loop adaptation, the resulting state ϕ'_t is evaluated on a small QA support set $\mathcal{D}_{\text{align}}$ to strengthen its use for downstream inference. For each $(q_i, a_i) \in \mathcal{D}_{\text{align}}$, the answer loss is computed with $\mathcal{R}_{q_i}$ retrieved from E to optimize joint memory use:

$$\mathcal{L}_{\text{outer}}(\phi) = -\frac{1}{|\mathcal{D}_{\text{align}}|} \sum_{(q_i,a_i) \in \mathcal{D}_{\text{align}}} \frac{1}{|a_i|} \sum_{j}^{|a_i|} \log p_\phi(a_{i,j} \mid q_i, \mathcal{R}_{q_i}, a_{i,<j}). \tag{4}$$

The corresponding gradient $\nabla_\phi \mathcal{L}_{\text{outer}}(\phi)$ is then used to update the initialization ϕ through a first-order approximation to improve inner-loop encoding:

$$\phi_{t+1} = \text{AdamW}_\eta \left(\phi_t,\ \text{ClipNorm} \left(\text{stopgrad} \left[\nabla_\phi \mathcal{L}_{\text{outer}}(\phi)|_{\phi=\phi'_t} \right], c \right) \right), \tag{5}$$

where ϕ_t denotes the initialization at iteration t and ϕ'_t encoded state after inner-loop adaptation. Here, η and c denote the outer-loop learning rate and gradient clipping threshold, respectively.

4 Adaptive Memory Composition during Inference

4.1 Diverse Memory Suitability

Although hybrid memory integrates the complementary properties of episodic and latent representations, different queries may favor different memory mechanisms, motivating adaptive modulation. As illustrated in Figure 4, uniform hybrid inference neglects query–memory compatibility and even can degrade answer quality relative to vanilla inference. When latent memory already captures the relevant information across past interactions, episodic retrieval may bias the inference toward instances integration at the expense of global coherence. Conversely, for queries answerable directly from episodic evidence, misaligned latent priors may lead the model to reinterpret the evidence to fit previously learned patterns. These findings suggest a reliable router for memory composition.

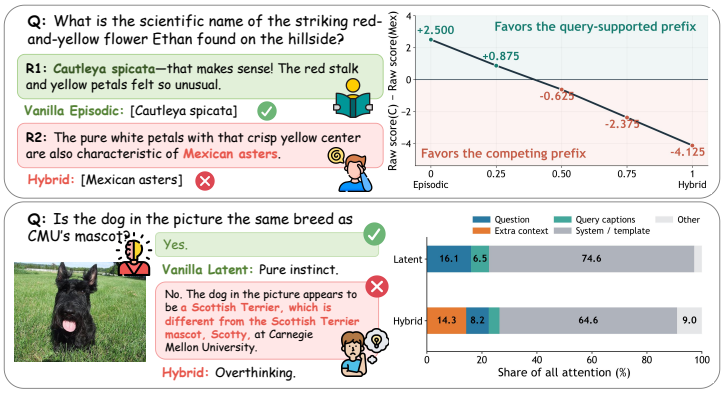


Figure 4: **Uniform hybrid configurations can degrade performance.** *Top*: Increasing hybrid strength draws episodic-locked evidence toward misleading context. *Bottom*: Extra evidence attention load disrupts coherent latent reasoning, leading to self-contradictions.

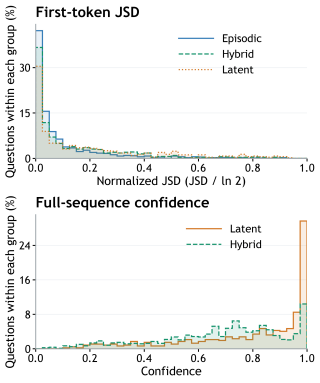


Figure 5: Distributions of first-token JSD and full-sequence confidence across winning samples for each memory paradigm.

4.2 Router Proxy

To efficiently adapt memory composition to the current query and memory state, a training-free, two-stage proxy mechanism is proposed. To assess how well internalized patterns align with the query, the first stage measures sensitivity to latent activation $d(q, \phi)$ by Jensen–Shannon divergence between the base and latent-only first-token distributions:

$$d(q, \phi) = JSD(P_\theta^{(1)}(q) || P_{\theta,\phi}^{(1)}(q)). \tag{6}$$

For queries retaining latent memory, the second stage further evaluates whether internalized knowledge is sufficient to answer the query using confidence $c_L(q, \phi)$:

$$c_L(q, \phi) = \exp\left(\frac{1}{T} \sum_{t=1}^{T} \log p_{\theta,\phi}\left(\hat{a}_t \mid q, \hat{a}_{<t} \right) \right). \tag{7}$$

Figure 5 compares the distributions of both proxies across queries grouped by successful memory modes. Although substantial overlap remains owing to the inclusion of co-winners, episodic-memory successes are more concentrated in the low-JSD region, supporting that vanilla episodic inference is more robust when latent responsiveness is weak. For confidence, latent winners exhibit a right-shifted distribution relative to hybrid winners with higher confidence for the current query.

4.3 Overall Framework

As Figure 3, given $\mathcal{R}$, construct $\mathcal{E}$ and ϕ following Section 3.2 and Section 3.3 respectively. For a query q, we first compute $d(q, \phi)$ as eq. (6) to determine whether to activate ϕ with calibrated τ_d:

$$g_L(q, \phi) = \mathbb{I}[d(q, \phi) \geq \tau_d]. \tag{8}$$

We further compute $c_L(q, \phi)$ via eq. (7) only when $g_L = 1$, and define the episodic gate using τ_c:

$$g_{\mathcal{E}}(q, \phi) = \mathbb{I}[d(q, \phi) < \tau_d \vee c_L(q, \phi) < \tau_c]. \tag{9}$$

When $g_{\mathcal{E}}(q, \phi) = 1$, we retrieve $\mathcal{R}_q$ according to $S(q, r)$ in eq. (2) and concatenate it with q. Final inference distribution is computed as follows, with $\mathcal{R}_q^{(0)} = \varnothing$ and $\mathcal{R}_q^{(1)} = \mathcal{R}_q$:

$$p(a \mid q) = p_{\theta,\, g_L(q,\phi)\cdot\phi}\left(a \mid q, \mathcal{R}_q^{(g_{\mathcal{E}}(q,\phi))}\right). \tag{10}$$

5 Experiments

5.1 Setups

Baselines and Benchmarks. We compare our approach against ten representative memory baselines, together with a no-memory setting: (a) **no external memory**: Vanilla Base, which performs direct inference with the backbone model; (b) **text-only retrieval-augmented methods**: Full Memory(Text), First-in-first-out(FIFO), NaiveRAG, Mem0 (Chhikara et al., 2025), and MemGPT (Packer et al., 2023); (c) **multimodal retrieval-augmented memory**: Full Memory(Multimodal), MuRAG (Chen et al., 2022), and UniversalRAG (Yeo et al., 2026); and (d) **latent memory**: MemVerse (Liu et al., 2025b), which combines long-term graph storage with a trained memory generator, while most alternatives require extensive training data. Evaluations are conducted on **Mem-Gallery** (Bei et al., 2026) and the **LoCoMo** (Maharana et al., 2024) subsets of MemoryBench (Ai et al., 2025), both targeting memory management within multi-turn multimodal conversations, simultaneously emphasizing fine-grained evidence recall and comprehensive reasoning.

Metrics. To evaluate memory performance, we employ both lexical and semantic metrics. Specifically, Token-level F1, BLEU-1, and Exact Match (EM) assess lexical answer quality (Rajpurkar et al., 2016; Papineni et al., 2002). For high-level semantic evaluation, we use DeepSeek-V4-Flash (DeepSeek-AI et al., 2026) as an LLM judge to score the correctness following Mem-Gallery protocols (Bei et al., 2026). and the LLM-as-a-Judge paradigm (Zheng et al., 2023).

Implementation Details. We conduct all experiments using Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the inference backbones. For methods requiring LLM-based memory construction, we adopt DeepSeek-V4-Flash (DeepSeek-AI et al., 2026). We implement latent memory using LoRA adapters with $r = 256$, applied to the final 4 decoder layers of each backbone. We reserve a small subset of the original test set for outer-loop training and evaluate all methods on the same remaining test samples, and Section 5.5 confirm that this protocol rules out train–test contamination. τ_d and τ_c are adaptively calibrated on the training set after the latent memory training phase. Further implementation details are provided in Appendix.

5.2 Main Results

Table 1 shows Hybrid and Adaptive achieve top scores on every metric across both benchmarks and model sizes. Compared to the strongest baseline, Hybrid improves Judge by 2.98%/2.24% on Mem-Gallery, and by 23.0%/12.5% on LoCoMo (for 3B/7B backbones). Beyond semantic correctness, the large gains in F1, BLEU-1, and EM indicate answering efficiency and stylistic alignment with memory coordination. Comparing individual memory modes reveals a strong complementarity between explicit evidence and implicit experience. For tasks emphasizing fine-grained details, sufficient evidence retrieval and effective organization within limited context windows are crucial: unimodal retrieval clearly underperforms multimodal retrieval without visual evidence. Our multi-index scoring episodic memory shows a strong foundation for evidence-grounded answering, performing competitively with MuRAG. While latent memory is less suited for recovering fine-grained

Table 1: **Main results on Mem-Gallery and LoCoMo.** Comprehensive evaluation of memory frameworks using Qwen2.5-VL (3B and 7B) across F1, BLEU-1, EM, and LLM-Judge. In addition to text-only and multimodal baselines, we report the Vanilla Base (no memory), standalone Episodic and Latent modules, fixed Hybrid composition, and Adaptive routing.

Method	Mem-Gallery				LoCoMo			
	F1	BLEU-1	EM	Judge	F1	BLEU-1	EM	Judge
Qwen-2.5-VL-3B								
Vanilla Base	29.69	26.03	18.21	36.14	29.36	26.16	10.00	34.00
FIFO	31.12	27.83	19.78	33.53	36.83	32.22	10.00	42.00
NaiveRAG	58.99	54.31	40.11	63.17	25.01	20.49	6.00	37.00
Mem0	29.89	26.67	19.96	31.98	36.94	33.02	10.00	46.00
Full-Context	39.05	35.84	26.16	40.17	35.01	31.55	16.00	50.00
MemGPT	39.06	35.84	26.16	41.13	37.01	33.55	18.00	43.00
Full-MM-Context	33.80	30.81	22.64	34.19	37.37	33.51	16.00	41.00
MuRAG	63.71	59.10	44.92	67.58	31.41	25.93	14.00	35.50
UniversalRAG	62.95	57.90	43.81	65.75	31.41	25.93	14.00	35.50
MemVerse	32.30	28.96	20.69	37.48	20.88	16.06	6.00	42.50
Episodic	62.90	57.96	43.16	67.28	38.30	32.30	16.00	42.00
Latent	38.87	34.95	25.69	38.54	34.10	20.52	4.00	48.50
Hybrid	69.38	64.68	52.50	70.56	67.42	63.37	**54.00**	73.00
Adaptive	**69.80**	**65.19**	**52.61**	**70.70**	**68.85**	**64.04**	**54.00**	**74.00**
Qwen-2.5-VL-7B								
Vanilla Base	29.28	25.55	15.90	36.23	44.15	39.50	12.00	48.00
FIFO	26.71	23.49	17.19	30.13	34.79	32.07	26.00	47.50
NaiveRAG	57.48	52.02	38.26	66.59	41.06	39.17	36.00	48.00
Mem0	30.24	26.52	18.30	35.00	43.50	41.90	38.00	50.00
Full-Context	35.72	32.21	24.58	40.32	38.73	35.08	26.00	53.50
MemGPT	52.69	47.88	32.81	66.10	42.31	37.20	20.00	56.50
Full-MM-Context	29.05	25.59	19.04	34.06	39.68	36.18	26.00	52.00
MuRAG	63.97	58.95	43.90	71.05	45.43	38.93	14.00	59.00
UniversalRAG	62.42	56.60	42.14	70.89	44.09	37.93	14.00	56.50
MemVerse	30.88	27.36	18.94	37.27	34.31	27.25	6.00	57.50
Episodic	61.89	56.54	42.33	70.61	48.12	43.29	30.00	63.50
Latent	37.74	33.63	25.51	40.02	33.40	19.80	4.00	50.00
Hybrid	71.27	66.76	**54.81**	73.29	**67.49**	**63.70**	**56.00**	71.50
Adaptive	**71.55**	**67.19**	54.76	**75.28**	65.56	61.10	52.00	**77.50**

details, it surpasses Full-(MM-)Context, proving that native internalization provides a more reliable representation than per-query in-context processing. MemVerse's LLM-based router assigns approximately 65% of queries to its latent memory branch, it still underperforms our standalone Latent. This degradation reveals the unreliability of internalizing hand-crafted memory storage with loss of global semantic information and highlights the gap between explicit external injection and native model inference utilization. Adaptive routing further improves performance (e.g., +1.2%/+2% Judge on Mem-Gallery/LoCoMo with 7B) by reliably selecting among the three memory modes.

5.3 Analysis for Task-Specific Memory Preferences

Following the task taxonomy established by Mem-Gallery (Bei et al., 2026), we analyze the performance trade-offs across Latent, Episodic, and Hybrid, as shown in Table 2, revealing distinct strengths and task-dependent benefits. For reasoning-intensive tasks including Timing Reasoning

Table 2: **Performance across task categories.** Comparison of Latent, Episodic, and Hybrid across nine task categories in Mem-Gallery with 7B, highlighting task-dependent strengths.

Method	FR	MR	TR	VR	TTL	VS	KR	CD	AR
Latent	21.10	33.97	37.97	26.80	62.38	2.08	38.94	59.62	**98.28**
Episodic	**67.38**	**73.09**	68.04	46.85	79.45	**73.18**	66.83	38.46	92.24
Hybrid	64.54	72.52	**69.94**	**52.03**	**84.25**	72.53	**69.71**	**69.23**	93.10

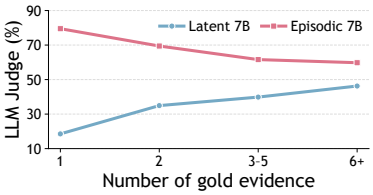


Figure 6: Performance under different gold evidence amounts.

(TR, +1.90%) and Visual Reasoning (VR, +5.18%), with Test-Time Learning (TTL, +4.81%) which demands rapid knowledge internalization, Hybrid and Latent show marked gains. These tasks benefit from compacting long-term visual facts to facilitate integrated reasoning. For Knowledge Resolution (KR, +2.88%) and Contradiction Detection (CD, +30.77%), which call for the comprehensive awareness, are substantially augmented by latent implication.

Conversely, for fine-grained, evidence-dependent tasks including Fact Retrieval (FR), Multi-Entity Reasoning (MR), and Visual Search (VS), pure Episodic maintains an advantage (+0.71%/+3.52%/+4.81%). These tasks rely heavily on grounded multimodal evidence without abstract loss, where Latent fall short (only 2.08% on VS and 21.10% on FR). For Answer Refusal (AR), which evaluates the ability to refrain from answering, Latent memory aids in recognizing broader context gaps, enabling more conservative refusal behaviors.

In addition, we evaluate performance relative to the number of required evidence in Figure 6. As gold evidence scales, Episodic degrades due to noise accumulation and retrieval precision loss. In contrast, Latent rises as inherently well-suited for global modeling. Overall, these task-dependent memory preferences highlight the need for memory integration and dynamic routing.

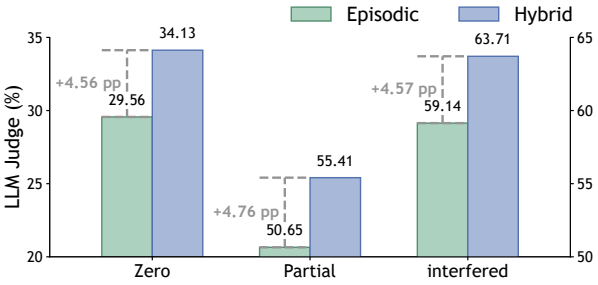


Figure 7: Comparison of Episodic and Hybrid under insufficient(zero and partial) and interfered retrieval.

Table 3: Training-label leakage analysis. *Seen(%)* is the fraction of test gold answers occurring in same-scene training labels. *Unseen@EM(%)* is the fraction of exact-match predictions whose gold answer is absent from the training labels.

Answer Interface	Seen	Unseen@EM
Open Semantic Answer	0.00	100.00
Local Evidence Pointer	15.64	95.04
Shared Canonical Label	52.86	37.16
Normalized Scalar	34.83	55.81

5.4 Latent Memory Compensates for Retrieval Failures

We examine latent memory as a complementary mechanism under primary Episodic failures caused by misaligned or incomplete retrievals. We compare Episodic with Hybrid under constructed imperfect-retrieval conditions: insufficient retrieval(zero and partial clue recall) and retrieval interference(wrong item ranks top-1), as Figure 7. Under partial and zero recall, Hybrid achieves relative gains of 9.5% and 15.2%, suggesting effective information completion by Latent. For retrieval interference, Hybrid improves Judge from 59.1% to 63.7%, highlighting its error correction capability. Together, these results demonstrate that latent memory effectively bridges the retrieval blind spots of hand-engineered strategies in long-tail scenarios.

5.5 Sensitivity to Hyperparameters

First, sweeping LoRA rank (Figure 8a) shows that while low-rank adapters yield initial gains, higher capacity better supports knowledge acquisition and joint use. Meanwhile, evaluating injection depth across the final n layers (Figure 8b) reveals peak accuracy at the top $n = 4$ layers, indicating adap-

Table 4: Average computational cost including Ingest latency per scene, QA latency and token usage per query.

Method	Latency (s)		Token Usage
	Ingest / scene	QA / q	Input tok / q
Mem0	1504.98	0.64	679.13
Full-Context	0.00	1.21	4569.82
Full-MM-Context	0.00	21.00	28 416.82
MuRAG	136.62	2.64	3274.82
UniversalRAG	138.77	3.82	2574.12
MemVerse	2663.50	19.94	2866.33
Latent	3244.72	0.68	510.84
Episodic	1958.61	10.91	2626.35

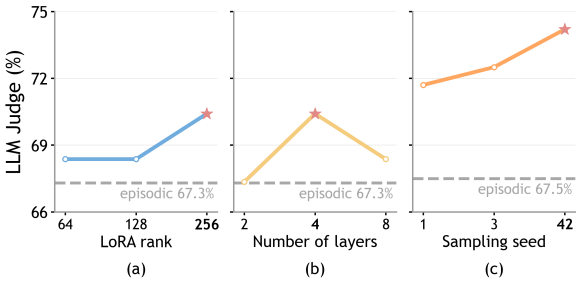


Figure 8: Hyperparameter sensitivity of Hybrid Memory on a Mem-Gallery scenario across varying (a) LoRA rank, (b) number of adapted layers, and (c) training-set sampling seed.

tation is best concentrated in layers that efficiently exploit latent memory while preserving core backbone features. Second, to rule out data selection sensitivity, we evaluate three resampled training sets on a fixed test set (Figure 8c), achieving consistent performance. Finally, to further exclude sole pattern memorization despite our conservative learning rate and limited optimization steps, we analyze train–test answer overlap across open-answer interfaces. According to Table 3, the two more open-ended and evidence-specific interfaces provide little opportunity for direct answer reuse. For the two more confounded interfaces, Shared Canonical Labels and Normalized Scalars, although they have a higher answer overlap, the most frequent same-scene training answer achieves only 8.45% and 11.24% EM, far below the Hybrid's 71.12% and 48.31%. Overall, these results confirm that latent memory indeed captures global implicit representations aligned with target outputs, rather than merely fitting training-set patterns.

5.6 Cost Analysis

Table 4 compares construction and query costs across methods on Mem-Gallery with Qwen2.5-VL-7B. While structured memory incurs upfront ingest overhead, it significantly optimizes test-time inference. Specifically, Episodic memory reduces input tokens by 42.5%, with pure inference taking $\sim$ 0.9s. Latent requires higher ingest latency due to meta-learning, but achieves extreme test-time efficiency (0.68s, 510.84 tokens) with superior performance over Full-Context. Overall, our approaches trade initial ingest overhead for optimal query efficiency and superior performance.

6 Conclusion

In this work, we characterize the complementary trade-offs of episodic and latent memory in multimodal agents and present a dual-level test-time adaptation framework. The framework couples a multi-indexed episodic memory layer for evidence grounding with a native latent adapter optimized via test-time adaptation for semantic internalization. To mitigate cross-branch interference, we further introduce a training-free adaptive routing mechanism driven by two proxies. Extensive experiments on multimodal long-term benchmarks demonstrate that the hybrid framework significantly outperforms single-paradigm baselines with superior robustness and efficiency. Moreover, different tasks show marked variations in their memory mechanism preferences across scenarios.

Limitation and Future works. **(1) Task Scenarios:** Constrained by available standard benchmarks that lean toward episodic memory, the framework can be extended to scenarios with higher procedural memory demands. **(2) Routing Strategy:** To minimize overhead, our router relies on low-cost proxies; lightweight trainable parameters can be integrated to potentially improve precision. **(3) Memory Evolution:** While both memories are constructed post-trajectory with fixed distributions, reflective collaborative updates can be incorporated to enable dynamic evolution.

7 Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

In this work, we used generative AI tools (such as large language models) to assist with language polishing, improving readability, and formatting references/formatting parts of the research manuscript. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims, or proving theoretical models/conceptual frameworks, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming, summarizing or analyzing existing literature, and debugging experimental code without writing. We have reviewed all AI-assisted work and take full responsibility for the final content of this work with the aid of generative AI.

8 Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

9 Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets will be made publicly available in an anonymous repository to facilitate replication and verification. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of our framework to assist others in reproducing our experiments. Additionally, Mem-Gallery and Memory Bench are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

References

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. Memorybench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*, 2025.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. *Advances in Neural Information Processing Systems*, 38:46968–47002, 2026a.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2026b.

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

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Zeming Chen, Angelika Romanou, Gail Weiss, and Antoine Bosselut. Perk: Long-context reasoning as parameter-efficient test-time learning. In *International Conference on Learning Representations*, volume 2026, pp. 10350–10377, 2026.

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

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Guhao Feng, Shengjie Luo, Kai Hua, Ge Zhang, Wenhao Huang, Di He, and Tianle Cai. In-place test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 114452–114472, 2026a.

Tao Feng, Chongrui Ye, Tianyang Luo, Jingjun Xu, Xueqiang Xu, Haozhen Zhang, Ge Liu, and Jiaxuan You. Elasticmem: Latent memory as a learnable resource for llm agents. *arXiv preprint arXiv:2605.30690*, 2026b.

Muxin Fu, Xiangyuan Xue, Yafu Li, Zefeng He, Siyuan Huang, Xiaoye Qu, Yu Cheng, and Yang Yang. Latentmem: Customizing latent memory for multi-agent systems. *arXiv preprint arXiv:2602.03036*, 2026.

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

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, et al. Evaluating long-horizon memory for multi-party collaborative dialogues. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9082–9090, 2026a.

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

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R Fung. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, volume 2025, pp. 20310–20334, 2025.

Dingyi Kang, Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. V-Mem: Modality-Routed Retrieval for Long-Term Multimodal Agentic Memory, 2026. URL https://arxiv.org/abs/2608.01543.

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, 2025a. URL https://arxiv.org/abs/2512.03627.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, et al. Memverse: Multimodal memory for lifelong learning agents. *arXiv preprint arXiv:2512.03627*, 2025b.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

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. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2actbench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8173–8190, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

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. *Advances in Neural Information Processing Systems*, 37:53764–53797, 2024.

Renxi Wang, Xudong Han, Lei Ji, Shu Wang, Timothy Baldwin, and Haonan Li. Toolgen: Unified tool retrieval and calling via generation. In *International Conference on Learning Representations*, volume 2025, pp. 73473–73498, 2025.

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

Zijun Wu, Yongchang Hao, and Lili Mou. Tokmem: One-token procedural memory for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 32460–32480, 2026.

Siyu Xia, Chenheng Zhang, Yanting Wu, Haoxuan Li, Jiajun Chai, Xiaohan Wang, Guojun Yin, Wei Lin, Zhouchen Lin, Haifeng Zhang, and Jun Wang. UniMem: Complementary Episodic-to-Parametric Memory for Boundary-Agnostic Task Streams, 2026. URL https://arxiv.org/abs/2607.26017.

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

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, 2026.

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31544–31555, 2026.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. In *International Conference on Learning Representations*, volume 2026, pp. 22555–22588, 2026a.

Zeyu Zhang, Rui Li, Xiaoyan Zhao, Yang Zhang, Wenjie Wang, Xu Chen, and Tat-Seng Chua. Nextmem: Towards latent factual memory for llm-based agents. *arXiv preprint arXiv:2603.15634*, 2026b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.