MORE MULTIMODALITY IS BETTER? UNCOVERING ITS FUNCTIONAL ROLES IN AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language model (MLLM) agents increasingly rely on multimodal memory to store and reuse past experiences for long-horizon tasks, yet it remains unclear when, where, and how multimodal information benefits memory. We present the first systematic study of how different levels of multimodal integration within the memory pipeline affect agent performance. We establish a four-level multimodal memory hierarchy that progressively introduces multimodal capabilities into memory construction, storage and retrieval, and utilization, ranging from text-only to fully multimodal memory. Experiments are conducted under long-horizon settings across 4 representative memory-intensive task domains, 10 benchmarks, and 4 MLLM backbones. Our results reveal that deeper multimodal integration does not consistently improve performance. Instead, its benefits depend on whether newly introduced multimodal capabilities address remaining evidence gaps. Three critical roles of multimodal information are identified: a Semantic Enricher during memory construction, an Experience Locator during retrieval, and an Evidence Resolver during utilization—capturing distinct ways multimodal information supports downstream decision-making. Deeper integration, however, may also introduce distraction and yield limited or negative returns. These findings challenge the prevailing assumption that richer multimodal memories universally benefit agent performance and highlight the importance of stage-aware multimodal memory design.

1 INTRODUCTION

As language model agents are increasingly deployed in complex environments that require long-horizon interaction, memory has become essential for retaining and reusing past experiences to support future decision making (Mohsin et al., 2025; Zhang et al., 2025b; Cai et al., 2025; Dou et al., 2026). While early agent memory systems primarily represented past experiences through textual abstractions (Ouyang et al., 2026; Zhang et al., 2025a; Suzgun et al., 2025; Wang et al., 2025), the growing prevalence of multimodal interactions has motivated memory systems to preserve and leverage information beyond text. Multimodal memory has consequently emerged as an important direction for agents operating in visually and multimodally rich environments (Zeng et al., 2026; Liu et al., 2026; Yeo et al., 2026; Wang et al., 2026a).

Broadly, an agent memory system can be viewed as a three-stage pipeline comprising memory construction, storage and retrieval, and memory utilization (Hu et al., 2025; Zhang et al., 2025c). Existing multimodal memory systems vary considerably in how deeply multimodal information is integrated across these stages (Jiang et al., 2026; Wang et al., 2026b; Yeo et al., 2026). However, it remains unclear what role multimodal information plays at each stage, whether deeper integration consistently improves performance, and under what conditions multimodal information provides value throughout the memory pipeline. To answer these questions, we conduct a systematic study of multimodal integration across the three stages of agent memory, organized around two core research questions (RQs).

We begin with a seemingly intuitive question: **(RQ1)** *Does increasing multimodal integration consistently improve memory performance?* To make "more multimodality" measurable, we con-

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.

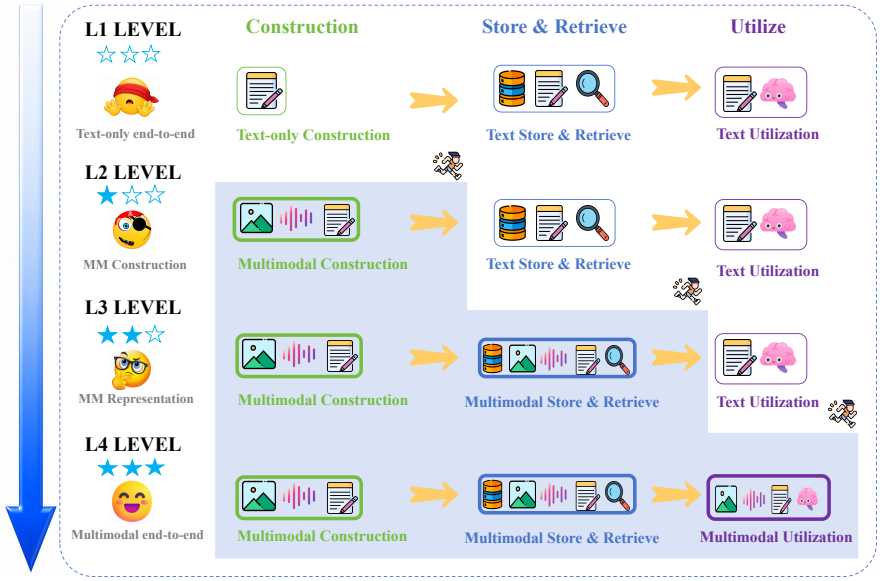


Figure 1: **Four-level multimodal memory hierarchy:** A structured hierarchy that progressively integrates multimodal information into different memory stages: memory construction, memory storage and retrieval, and memory utilization.

struct a four-level hierarchy that progressively extends multimodal information through the memory pipeline. As illustrated in Figure 1, the hierarchy ranges from text-only memory (L1) to multimodal integration in memory construction (L2), storage and retrieval (L3), and utilization for downstream decision-making (L4). This turns multimodal integration into a controlled experimental axis and provides a unified view of existing systems. We evaluate the hierarchy across 10 benchmarks spanning 4 representative long-horizon multimodal scenarios, using controlled L1–L4 variants over representative memory frameworks and multiple MLLM backbones.

The answer to RQ1 is clear: more multimodality is not consistently better. Performance does not increase monotonically from L1 to L4. Instead, different scenarios favor different integration depths (§4.1). This pattern persists across diverse benchmarks, different backbones, and alternative memory implementations (§4.2). Stronger models improve overall performance, but do not produce a universally optimal integration level. These results suggest that the key question is not simply *how much* multimodal information memory retains, but *what it is needed for* at different stages of the pipeline.

This naturally leads to our second question: **(RQ2)** *How can multimodality provide value in the agent memory pipeline?* We first show that aggregate benchmark performance can be misleading for evaluating multimodal memory, as many benchmarks mix genuinely multimodal instances with queries that remain answerable from text alone (§5.1). We then show that memory orientation provides an informative distinction: in transfer-oriented settings, retrieved historical experience often offers limited incremental relevance beyond the current task context, constraining the benefit of richer multimodal memory (§5.2). For evidence-oriented settings, we develop an evidence-gap perspective and identify three stage-specific roles of multimodal information—Semantic Enricher, Experience Locator, and Evidence Resolver—corresponding to construction, retrieval, and utilization. We further show that multimodal benefits are stage-local rather than depth-monotonic: additional capability helps when it resolves a remaining stage-specific evidence gap, but can yield limited or negative returns otherwise (§5.3).

In conclusion, our findings shift multimodal memory design from a modality-centric view of "more is better" toward a function-aware perspective. The value of multimodality depends on whether it is genuinely needed, whether the memory provides relevant information, and whether the added capability addresses an unresolved evidence need at the appropriate stage. Rather than maximizing

multimodal integration, effective memory systems should align multimodal capability with stage-specific information needs—better multimodal memory may come not from remembering more modalities, but from knowing *where* and *why* each modality matters.

2 PRELIMINARIES

We formulate an agent memory system as a three-stage pipeline: memory construction, storage and retrieval, and memory utilization.

Memory Construction. Given a raw memory source $U = \{u_t\}_{t=1}^{T}$, the construction module extracts reusable memory items

$$M = f_\theta^c(U), \tag{1}$$

where $M = \{m_i\}_{i=1}^{N}$ denotes the resulting memory bank.

Memory Storage and Retrieval. Given a query q, the retrieval module selects top-k relevant memories from M:

$$R(q) = \mathrm{TopK}_{m_i \in M}\, s_\theta^r(q, m_i), \tag{2}$$

where s_θ^r denotes the retrieval score.

Memory Utilization. The retrieved memories are incorporated with the current task input x to produce the final output:

$$y = f_\theta^u(x, R(q)). \tag{3}$$

This three-stage abstraction provides a common interface for analyzing where multimodal information is introduced into the memory pipeline.

3 CONTROLLED MULTIMODAL INTEGRATION IN AGENT MEMORY

Building on the three-stage memory pipeline defined in Section 2, we construct a controlled hierarchy that progressively introduces multimodal processing into memory construction, storage and retrieval, and utilization. This design isolates the effect of multimodal information at different stages while keeping the underlying memory framework fixed.

3.1 MULTIMODAL INTEGRATION HIERARCHY

We represent the multimodal configuration of memory level ℓ using a three-dimensional gate vector

$$\mathbf{g}^{(\ell)} = \left[g_c^{(\ell)}, g_r^{(\ell)}, g_u^{(\ell)}\right] \in \{0,1\}^3, \tag{4}$$

where g_c, g_r, and g_u indicate whether multimodal information is enabled during memory construction, storage and retrieval, and utilization, respectively.

We define four cumulative levels of multimodal integration:

$$L_1 : [0,0,0], \quad L_2 : [1,0,0], \quad L_3 : [1,1,0], \quad L_4 : [1,1,1]. \tag{5}$$

L_1: Text-only Memory. All stages operate on textual representations. Multimodal observations are excluded from the memory pipeline, providing the text-only baseline.

L_2: Multimodal Construction. Multimodal inputs are directly used during memory construction, allowing visual information to influence what is extracted into memory. The resulting memory representation remains textual, and subsequent retrieval and utilization operate on text.

L_3: Multimodal Storage and Retrieval. In addition to multimodal construction, multimodal representations are preserved for storage and retrieval. Retrieval can therefore exploit both textual and visual information, while the retrieved memory provided to downstream reasoning remains textual.

L_4**: Fully Multimodal Memory.** Multimodal information is retained throughout the full memory pipeline. Retrieved visual evidence is directly exposed to the MLLM during memory utilization and final decision making.

The four levels therefore form a cumulative hierarchy:

$$L_1 \xrightarrow{+\text{MM Construction}} L_2 \xrightarrow{+\text{MM Storage \& Retrieval}} L_3 \xrightarrow{+\text{MM Utilization}} L_4. \tag{6}$$

This cumulative design enables adjacent-level comparisons to isolate the incremental contribution of multimodal information at each memory stage.

3.2 Experimental Instantiation

We instantiate all 4 multimodal memory levels within the same base memory framework for each task domain, introducing multimodal capability only at the designated memory stage while keeping other experimental settings as consistent as possible.

Memory Frameworks. We use ReasoningBank (Ouyang et al., 2026) for GUI interaction and visual reasoning, Omni-SimpleMem (Liu et al., 2026) for conversation understanding, and WorldMM (Yeo et al., 2026) for video understanding. These are mainstream memory frameworks in their domains. (details in Appendix A)

Controlled Instantiation. For each framework, we construct L_1–L_4 variants from the same source experiences and preserve the original memory organization and inference pipeline as much as possible. Across levels, we control the task setting, prompting, evaluation procedure, and downstream memory-evidence budget fixed while changing only the modality available at construction, storage and retrieval, or utilization. Framework-specific implementations are detailed in Appendix C.2.

Benchmarks. We evaluate on 10 benchmarks across 4 domains: WebVoyager (He et al., 2024) and Mind2Web (Deng et al., 2023) for GUI interaction; AgentVista (Su et al., 2026) and MMSearch-Plus (Tao et al., 2025) for visual reasoning; Mem-Gallery (Bei et al., 2026) and MMRC (Xue et al., 2025), with MemLens (Ren et al., 2026) and MemEye (Guo et al., 2026) for additional analysis, for conversation understanding; and Video-MME (Fu et al., 2025) and HippoVlog (Lin et al., 2026) for video understanding. Together, these benchmarks cover diverse multimodal agent memory applications (details in Appendix B).

MLLM Backbones. Experiments use 4 MLLM backbones: Qwen3-VL-32B-Instruct (Bai et al., 2025), Qwen3.5-27B (Qwen Team, 2026), GPT-5.4-nano (Singh et al., 2025), and Gemini-2.5-Flash (Comanici et al., 2025), covering different model families, scales, and deployment settings.

4 Evaluating Multimodal Integration Depth

4.1 Performance across Multimodal Integration Levels

We first examine whether increasing multimodal integration consistently improves memory performance. Figure 2 compares L_1–L_4 across 8 benchmarks from 4 task domains using Qwen3-VL-32B-Instruct and GPT-5.4-nano. Other backbone results are provided in Appendix D.1.

Increasing multimodal integration yields non-monotonic returns. Across benchmarks and backbones, progressively extending multimodal information to additional stages of the memory pipeline does not produce a stable positive marginal gain. Deeper integration can substantially improve performance on some benchmarks, while providing only marginal gains or even degrading performance on others. For example, video understanding generally benefits from deeper multimodal integration, whereas several GUI interaction and visual reasoning benchmarks favor shallower or intermediate levels. Thus, integration depth itself is not a reliable predictor of memory performance.

Preferred multimodal memory integration depth varies across memory scenarios. The preferred integration level also differs substantially across benchmarks. WebVoyager and MMRC favor shallow integration, Mind2Web and MMSearch-Plus prefer an intermediate depth of L_2, Mem-Gallery consistently favors L_3, while VideoMME and HippoVlog benefit most from deeper multimodal integration. Moreover, the preferred level can vary with the backbone, as observed on AgentVista. These results indicate that there is no universally optimal multimodal integration depth.

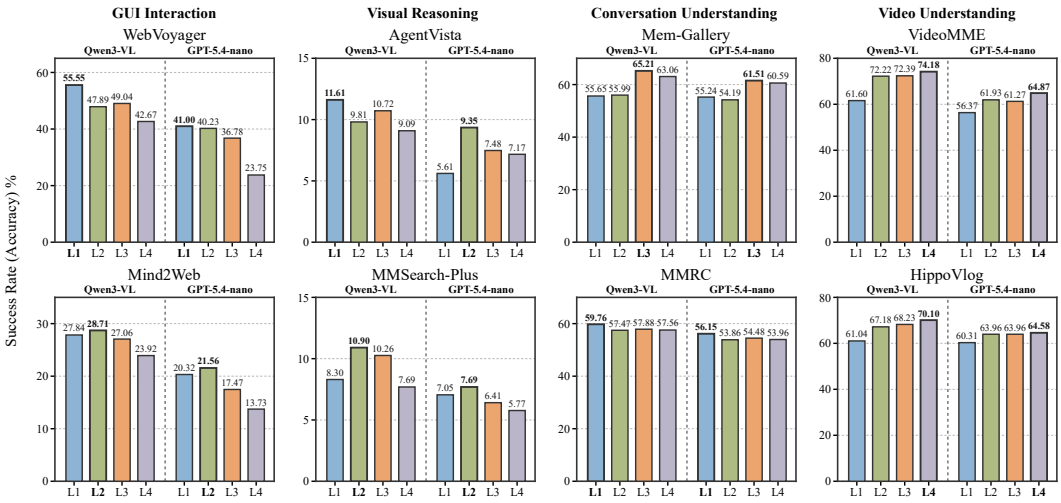


Figure 2: Performance across four multimodal memory integration levels on 8 benchmarks using Qwen3-VL-32B-Instruct (Qwen3-VL) and GPT-5.4-nano. Deeper multimodal integration does not consistently improve performance, and the preferred integration depth varies across benchmarks.

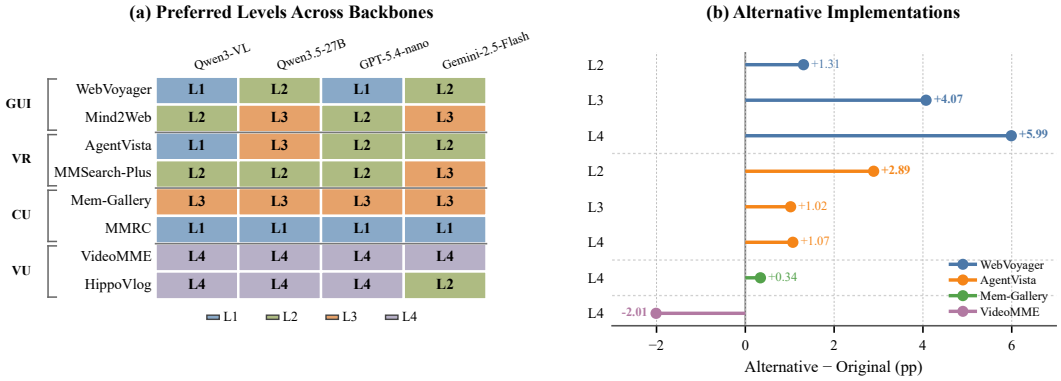


Figure 3: Robustness of multimodal integration patterns across MLLM backbones and memory implementations. (a) Preferred memory levels across 4 backbones on 8 benchmarks. (b) Performance changes under alternative implementations relative to the original ones.

This heterogeneity suggests that the effectiveness of multimodal memory is governed not by integration depth alone, motivating a deeper examination of when and where multimodality provides value in the memory pipeline in Section 5.

4.2 Robustness across Backbones and Implementations

We next examine whether the non-monotonic effect observed above remains robust across MLLM backbones and memory implementations.

MLLM backbones affect preferred integration depth. As shown in Figure 2, Qwen3-VL-32B-Instruct performs universally better than GPT-5.4-nano, while Figure 3 (a) summarizes the preferred integration levels across all 4 backbones, including Qwen3.5-27B and Gemini-2.5-Flash. Some benchmarks retain similar preferences across models, whereas others shift to different integration levels. Nevertheless, deeper multimodal integration does not become consistently preferable under any backbone.

Memory implementations substantially affect absolute performance. To assess whether our conclusions depend on a particular implementation choice, we construct alternative realizations for selected memory levels while keeping the level definitions and other experimental settings un-

changed. For GUI interaction and visual reasoning, we vary L_2–L_4: the L_2 alternative replaces direct access to the most recent k images with an MMSkill-based pipeline (Zhang et al., 2026), and the L_3 alternative replaces mandatory pre-task retrieval with model-controlled, on-demand memory access (Zhao et al., 2026). For conversation and video understanding, where visual and textual contents are intrinsically aligned and retrieved memories already serve as the primary answer evidence, we vary only L_4, replacing direct access to retrieved images with textual captions before reasoning. Full implementation details are provided in Appendix C. As shown in Figure 3(b), these alternative implementations can noticeably affect performance. Several configurations on WebVoyager and AgentVista improve over the original implementations, while the effect on Mem-Gallery is small and VideoMME shows a decrease. Nevertheless, changing the concrete realization of a memory level does not establish a consistent advantage for deeper multimodal integration: higher levels still fail to systematically outperform shallower ones.

Overall, our experiments consistently show that increasing multimodal integration does not guarantee better memory performance. This pattern remains robust across diverse benchmarks, MLLM backbones, and memory implementations, indicating that the preferred integration depth is not universal but varies with the underlying memory requirements.

5 Understanding the Uneven Returns to Multimodal Memory

Section 4 shows that deeper multimodal integration does not consistently improve performance, and that the best-performing memory level varies across evaluation settings. These findings motivate our **RQ2**: *How can multimodality provide value in the agent memory pipeline?* We first examine whether multimodal information is required for every instance. We then distinguish transfer-oriented from evidence-oriented memory according to how stored information contributes to the current task. Finally, for evidence-oriented settings, we analyze stage-specific evidence gaps to identify where multimodal capabilities provide value in the memory pipeline.

5.1 Overall Performance Can Obscure Multimodal Need

Although many benchmarks are designed as multimodal, multimodal dependence is often sparse at the instance level: only a subset of queries requires visual evidence from prior turns, while many remain answerable from textual history alone. This issue is particularly evident in conversation understanding and has also been noted in recent work (Guo et al., 2026). To examine this effect, we reevaluate 4 conversation-understanding benchmarks on their multimodal-only subsets using the same Qwen3-VL-32B-Instruct backbone and L_1–L_4 memory configurations. Detailed sample compositions are in Appendix D.3. In addition to comparing the preferred memory level, we measure how the relative gain of each multimodal level over the text-only baseline changes after filtering:

$$G_k = \mathrm{Perf}(L_k) - \mathrm{Perf}(L_1), \qquad \Delta G_k = G_k^{\mathrm{MM}} - G_k^{\mathrm{Full}},$$

where a positive ΔG_k indicates that the benefit of multimodal memory becomes larger on genuine multimodal subsets.

As shown in Figure 4, restricting evaluation to multimodal-only instances reveals substantially stronger multimodal gains on several benchmarks. Most notably, MMRC shifts from preferring L_1 on the full benchmark to L_4 on its multimodal subset. Even when the preferred level remains unchanged, the relative advantage of multimodal memory can increase substantially, as observed on Mem-Gallery and MemLens.

These results suggest that weak gains on the full benchmark do not necessarily imply that multimodal memory is ineffective. Aggregate evaluation can underestimate the utility of multimodal information when its demand is sparse at the instance level.

5.2 Transfer-Oriented Memory Often Offers Limited Incremental Relevance

Beyond multimodal necessity, we examine how stored history relates to the information required by the current task.

Transfer-Oriented versus Evidence-Oriented Memory. In our evaluation settings, GUI interaction and visual reasoning use *transfer-oriented memory*, where experience from previous tasks is

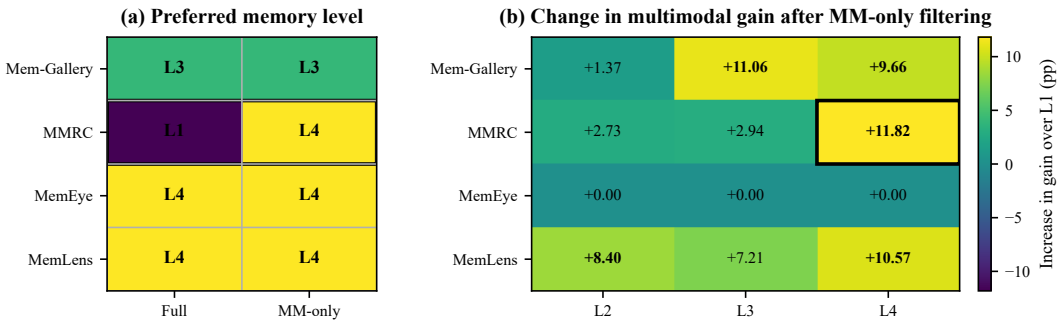


Figure 4: Effect of multimodal sample composition on memory evaluation. (a) Preferred memory levels on the full benchmarks and their multimodal-only subsets. (b) Increase in the gain of L_2–L_4 over the text-only L_1 baseline after restricting evaluation to multimodal-only samples. Positive values indicate that multimodal memory provides larger relative gains on genuinely multimodal instances.

Table 1: Incremental relevance of retrieved historical experiences beyond the current-task context. Mean Relevance averages the 0–3 relevance scores, while Actionable Relevance reports the percentage of retrieved histories with scores of at least 2.

	GUI Interaction		Visual Reasoning	
	WebVoyager	Mind2Web	AgentVista	MMSearch-Plus
Mean Relevance (0–3)	0.40	1.03	0.75	0.62
Actionable Relevance (%)	5.7	31.0	26.0	19.1

retrieved to guide a new task and is useful only when it provides transferable knowledge beyond the current context (Ouyang et al., 2026). By contrast, conversation and video understanding use *evidence-oriented memory*, where memory preserves the source content being queried and serves as answer evidence (Yeo et al., 2026). The distinction is therefore whether memory transfers prior experience across tasks or preserves source evidence for the current query.

Evidence from Retrieval Relevance. We audit the L_1 retrieval results on WebVoyager, Mind2Web, AgentVista, and MMSearch-Plus using an LLM judge that scores the incremental relevance of retrieved historical trajectories to the current task on a 0–3 scale, with scores of at least 2 treated as actionable. We report mean relevance and actionable relevance, with full protocols provided in Appendix E. As shown in Table 1, mean relevance ranges from 0.40 to 1.03 and actionable relevance from 5.7% to 31.0%, indicating that most retrieved histories provide limited task-specific guidance beyond the current context. Another consistent performance pattern further supports this observation: across GUI interaction and visual reasoning benchmarks, moving from L_3 to L_4 consistently degrades performance across backbones.

Together, these results suggest that the applicability of past experience is a key bottleneck in transfer-oriented settings: preserving richer multimodal details is unlikely to help when the underlying experience itself is only weakly relevant, and may instead introduce distracting historical content.

5.3 Multimodal Benefits Are Stage-Local

In evidence-oriented settings, memory represents the source content being queried, yet answer-relevant information may still be omitted during construction, missed during retrieval, or lost when visual content is compressed into text. We refer to such missing information as an *evidence gap*. Our central hypothesis is that gains from multimodal memory depend on whether visual information addresses the evidence gap remaining at each stage.

Table 2 summarizes three functional roles—Semantic Enricher, Experience Locator, and Evidence Resolver—and their expected signatures. To evaluate these roles, we conduct a paired evidence analysis on VideoMME, HippoVlog, Mem-Gallery, and MMRC.

Table 2: Stage-specific roles and their testable predictions.

Transition	Role	Residual Bottleneck	Expected Effect
L1→ L2	Semantic Enricher	Visual semantics are missing from textual memory.	L2 enriches textual memory with answer-relevant visual semantics.
L2→ L3	Experience Locator	The answer-bearing memory is missed by text-based retrieval.	L3 locates answer-bearing memories by multimodal retrieval.
L3→ L4	Evidence Resolver	Retrieved textual memory lacks discriminative visual detail.	L4 provides visual details that are lost in textual abstraction.

Table 3: Evidence and performance changes across multimodal memory transitions. Answer Support ranges from 0 to 3; sufficiency, Rescue, Loss, and Acc. are percentages; ΔAcc. is in percentage points. All results refer to the corresponding paired audit set.

	Baseline	Evidence Change			Task Effect
Benchmark	**Sufficiency**	**Answer Support**	**Rescue**	**Loss**	**Baseline Acc. (ΔAcc.)**
Semantic Enricher	L1→L2				
VideoMME	43.8	1.97 → 2.72	37.5	2.1	61.60 ↑(+10.62)
HippoVlog	48.5	2.08 → 2.51	26.8	6.6	61.04 ↑(+6.14)
Mem-Gallery	56.0	2.20 → 2.22	9.2	7.5	55.65 ↑(+0.34)
MMRC	71.0	2.45 → 2.43	4.12	5.70	59.76 ↓(−2.29)
Experience Locator	L2→L3				
VideoMME	81.5	2.74 → 2.74	1.1	1.5	72.22 ↑(+0.17)
HippoVlog	75.1	2.64 → 2.62	4.8	5.7	67.18 ↑(+1.05)
Mem-Gallery	57.7	2.19 → 2.33	7.5	1.2	55.99 ↑(+9.22)
MMRC	69.1	2.43 → 2.45	1.99	0.56	57.47 ↑(+0.41)
Evidence Resolver	L3→L4				
VideoMME	88.3	2.83 → 2.86	1.9	0.2	72.39 ↑(+1.79)
HippoVlog	71.5	2.55 → 2.65	5.4	0.9	68.23 ↑(+1.87)
Mem-Gallery	76.8	2.59 → 2.58	0.2	0.7	65.21 ↓(−2.15)
MMRC	67.7	2.40 → 2.42	0.97	0.05	57.88 ↓(−0.32)

Notes. Baseline sufficiency: $S^- = 3$; Rescue: $S^- < 3$, $S^+ = 3$; Loss: $S^- = 3$, $S^+ < 3$. Sufficiency, Rescue, Loss, and Acc. are reported as percentages; ΔAcc. denotes the accuracy change in percentage points. Sampling and coverage details are provided in the appendix.

For each query, an LLM judge jointly evaluates the evidence from two adjacent memory levels against the same question and reference answer, assigning support scores S^- and S^+ ranging from 0-3 to the lower- and higher-level conditions, respectively. Table 3 reports mean Answer Support together with Baseline Sufficiency, Evidence Rescue, and Evidence Loss, capturing whether added multimodal capabilities close or introduce evidence gaps. These metrics assess evidence availability rather than final answer correctness; formal definitions and evaluation details are provided in Appendix F. We next examine how each transition addresses a distinct stage-specific evidence gap.

Semantic Enricher ($L_1 \rightarrow L_2$). Multimodal construction can encode visual semantics absent from text-only inputs into textual memory. The video benchmarks provide the clearest evidence: Answer Support rises from 1.97 to 2.72 on VideoMME and from 2.08 to 2.51 on HippoVlog, accompanied by performance gains of 10.62 and 6.14 points. This supports a construction-stage role in which visual information strengthens the answer-bearing semantics preserved in memory.

Experience Locator ($L_2 \rightarrow L_3$). Multimodal retrieval can recover answer-bearing memories missed by text-based matching. Mem-Gallery provides the clearest example: Answer Support increases from 2.19 to 2.33, Rescue exceeds Loss (7.5% vs. 1.2%), and performance improves by 9.22 points. Here, visual information primarily determines which memory is retrieved, while the evidence used for downstream reasoning remains textual.

Evidence Resolver ($L_3 \rightarrow L_4$). When retrieved textual memory lacks discriminative visual detail, direct access to raw evidence can resolve the remaining ambiguity. Both video benchmarks show this pattern: Rescue exceeds Loss and performance further improves at L_4, supporting the value of visual evidence beyond textual descriptions. Unlike semantic enrichment, this role preserves original visual detail until the final reasoning stage.

The positive cases above characterize when multimodal information closes a residual evidence gap, but additional capability does not guarantee additional gain. Mem-Gallery illustrates this stage locality most clearly: $L_1 \rightarrow L_2$ yields only +0.34 points, $L_2 \rightarrow L_3$ yields +9.22 points, whereas $L_3 \rightarrow L_4$ decreases performance by 2.15 points with almost no Evidence Rescue (0.2%). MMRC shows a similar negative return at $L_1 \rightarrow L_2$, where Evidence Loss exceeds Rescue. These cases are consistent with an evidence–burden trade-off: when newly introduced multimodal capability contributes little additional evidence, the added multimodal context can offset its benefit.

Together, these roles show that deeper multimodal integration introduces distinct capabilities across the memory pipeline, but their value is conditional on the bottleneck they address. Visual information improves performance only when the newly introduced capability closes an unresolved evidence gap. **Effective multimodal memory design is not about maximizing multimodal capability, but about aligning it with stage-specific evidence needs.**

6 Related Work

Memory-Augmented Agents. Existing memory-augmented agents can be broadly divided into two paradigms according to the role of memory. *Experience-based memory* stores or distills past interactions into executable workflows, reasoning patterns, or skills that provide complementary guidance for future tasks (Wang et al., 2025; Ouyang et al., 2026; Mi et al., 2026). In contrast, *context-preserving memory* retains information from long-form sources such as conversations, videos, and documents (Zhong et al., 2024; Long et al., 2026; Yang et al., 2026), where retrieved memories often serve as direct evidence for downstream reasoning.

Multimodal Memory Systems. Recent studies extend text-based memory by introducing multimodal information at different stages of the memory pipeline. Following the four-level integration hierarchy defined in Section 3.1, we categorize prior systems by their multimodal integration depth, as summarized in Appendix Table 6. Some methods use multimodal inputs only during memory construction and store textual abstractions (Jiang et al., 2026; Zuo et al., 2026). Others preserve multimodal representations for indexing and cross-modal retrieval while still providing textual memory to downstream reasoning (Wang et al., 2026b). More recent systems directly expose retrieved multimodal evidence during memory utilization (Yeo et al., 2026; Liu et al., 2026).

Despite this progress, existing methods typically introduce multimodality together with changes in memory representations and pipelines, making it difficult to isolate the contribution of multimodal information itself. We instead adopt a unified pipeline-level framework that systematically controls multimodal integration across construction, storage and retrieval, and utilization, enabling direct comparison of different integration depths.

7 Conclusion

In this work, we present the first systematic study of how multimodal integration across the memory pipeline affects agent performance. By establishing a four-level multimodal memory hierarchy and controlled memory designs, we show that deeper integration does not consistently improve performance. Our deeper analysis reveals two key bottlenecks: multimodal gains can be diluted by text-only samples, while transfer-oriented memory is constrained by experience relevance. For evidence-oriented settings, we identify three stage-specific roles—Semantic Enricher, Experience Locator, and Evidence Resolver—and show that multimodality helps only when it closes a residual evidence gap. These findings suggest a shift from maximizing multimodal integration depth to matching multimodal capability to the actual needs of the memory process, offering insights for effective multimodal memory design.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and polishing, including improvements to clarity, wording, grammar, organization, and presentation. LLMs were also used to support parts of the research workflow, including LLM-based evaluation, analysis and interpretation of experimental results, and refinement of scientific figures and their presentation. All AI-assisted analyses, figures, and written contents were reviewed and verified by the authors against the underlying experimental results and source materials. The authors take full responsibility for the final text, claims, and artifacts presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the paper and appendix. Section 3.2 describes the controlled experimental setup and benchmark configuration. Appendix B provides dataset statistics, subset selection, and evaluation settings, while Appendix C details model parameters, retrieval settings, and framework-specific memory implementations. The protocols and metric definitions for the incremental relevance audit and paired answer-support audit are provided in Appendices E and F, respectively. These details are intended to enable reproduction of the experimental setup, evaluation procedures, and analysis reported in this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

Yuxuan Cai, Yipeng Hao, Jie Zhou, Hang Yan, Zhikai Lei, Rui Zhen, Zhenhua Han, Yutao Yang, Junsong Li, Qianjun Pan, et al. Building self-evolving agents via experience-driven lifelong learning: A framework and benchmark. *arXiv preprint arXiv:2508.19005*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Shihan Dou, Ming Zhang, Chenhao Huang, Jiayi Chen, Feng Chen, Shichun Liu, Yan Liu, Chenxiao Liu, Cheng Zhong, Zongzhang Zhang, et al. Evalearn: quantifying the learning capability and efficiency of llms via sequential problem solving. *Advances in Neural Information Processing Systems*, 38:125892–125945, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025.

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.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Helen Li, and Yiran Chen. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5977, 2026.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, et al. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory. *arXiv preprint arXiv:2604.01007*, 2026.

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.

Qirui Mi, Zhijian Ma, Mengyue Yang, Haoxuan Li, Yisen Wang, Haifeng Zhang, and Jun Wang. Skill-pro: Learning reusable skills from experience via non-parametric ppo for llm agents. *arXiv preprint arXiv:2602.01869*, 2026.

Muhammad Ahmed Mohsin, Muhammad Umer, Ahsan Bilal, Zeeshan Memon, Muhammad Ibtsaam Qadir, Sagnik Bhattacharya, Hassan Rizwan, Abhiram R Gorle, Maahe Zehra Kazmi, Ayesha Mohsin, et al. On the fundamental limits of llms at scale. *arXiv preprint arXiv:2511.12869*, 2025.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jL7fwchScm.

Qwen Team. Qwen3.5-27b. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, et al. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

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

Zhaochen Su, Jincheng Gao, Hangyu Guo, Zhenhua Liu, Lueyang Zhang, Xinyu Geng, Shijue Huang, Peng Xia, Guanyu Jiang, Cheng Wang, et al. Agentvista: Evaluating multimodal agents in ultra-challenging realistic visual scenarios. *arXiv preprint arXiv:2602.23166*, 2026.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*, 2025.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*, 2025.

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*, 2026a.

Shijian Wang, Jiarui Jin, Runhao Fu, Zexuan Yan, Xingjian Wang, Mengkang Hu, Eric Wang, Xiaoxi Li, Kangning Zhang, Li Yao, et al. Museagent: A multimodal reasoning agent with stateful experiences. *arXiv preprint arXiv:2603.27813*, 2026b.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Forty-second International Conference on Machine Learning*, 2025.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22477–22503, 2025.

Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents. *arXiv preprint arXiv:2603.03296*, 2026.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Ziyun Zeng, Hang Hua, Bocheng Zou, Mu Cai, Rogerio Feris, and Jiebo Luo. Mementogui: Learning agentic multimodal memory control for long-horizon gui agents. *arXiv preprint arXiv:2605.18652*, 2026.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025a.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. Memevolve: Meta-evolution of agent memory systems. *arXiv preprint arXiv:2512.18746*, 2025b.

Kangning Zhang, Shuai Shao, Qingyao Li, Jianghao Lin, Lingyue Fu, Shijian Wang, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, et al. Mmskills: Towards multimodal skills for general visual agents. *arXiv preprint arXiv:2605.13527*, 2026.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025c.

Weixiang Zhao, Yingshuo Wang, Yichen Zhang, Yanyan Zhao, Yu Zhang, Yang Wu, Dandan Tu, Bing Qin, and Ting Liu. Rethinking experience utilization in self-evolving language model agents. *arXiv preprint arXiv:2605.07164*, 2026.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *Advances in Neural Information Processing Systems*, 38:25660–25691, 2026.

A Agent Frameworks

We provide detailed descriptions of the three representative memory frameworks used as the underlying agent architectures in our experiments: ReasoningBank (Ouyang et al., 2026), Omni-SimpleMem (Liu et al., 2026), and WorldMM (Yeo et al., 2026). We select these frameworks because they represent strong and representative memory designs for the corresponding task domains, covering experience transfer across tasks, long-term multimodal conversation memory, and video understanding. Using established domain-specific frameworks allows us to study multimodal integration under realistic memory pipelines rather than relying on a single customized memory architecture. Specifically, ReasoningBank is adopted for GUI interaction and visual reasoning, Omni-SimpleMem for conversation understanding, and WorldMM for video understanding.

ReasoningBank. ReasoningBank is an experience-driven memory framework that enables agents to improve over a sequence of tasks by distilling reusable reasoning strategies from past interactions. After completing a task, the agent evaluates the resulting trajectory and extracts generalizable reasoning patterns from both successful and failed experiences. These distilled reasoning memories are accumulated in an external memory bank. When solving a new task, the agent retrieves memories that are relevant to the current problem and incorporates them into its context as guidance for subsequent reasoning and action. Newly acquired experiences can in turn be distilled and added back to the memory bank, forming a continual experience–retrieval–update loop without modifying the parameters of the underlying language model. In our experiments, ReasoningBank serves as the base memory framework for GUI interaction and visual reasoning tasks.

Omni-SimpleMem. Omni-SimpleMem is a lifelong multimodal memory framework designed to store and retrieve information from extended multimodal interaction histories. It supports multiple modalities, including text, images, audio, and video, and organizes them into a unified external memory for later query-based access. The framework selectively retains informative content during memory construction, combines semantic and lexical signals for progressive retrieval, and further augments retrieved information with structured cross-modal relations to support more complex reasoning. At inference time, memories relevant to the current query are retrieved and provided as contextual evidence for downstream response generation. In our experiments, Omni-SimpleMem serves as the base memory framework for conversation-understanding tasks.

WorldMM. WorldMM is a multimodal memory framework designed for long-video understanding and reasoning. It constructs three complementary memory components from video streams: episodic memory that represents events across multiple temporal scales, semantic memory that maintains high-level and long-term relational knowledge, and visual memory that preserves fine-grained visual details that may be lost in textual abstraction. At inference time, an adaptive retrieval agent iteratively selects relevant information from different memory sources and temporal granularities according to the current query. The retrieved evidence, together with the retrieval history, is then provided to a response agent for final answer generation. In our experiments, WorldMM serves as the base memory framework for video-understanding tasks.

B Environments and Benchmarks

We evaluate our framework on a diverse suite of 10 representative benchmarks spanning four major multimodal agent memory scenarios: *GUI interaction*, *visual reasoning*, *conversation understanding*, and *video understanding*. The benchmark suite covers substantially different interaction formats, memory requirements, temporal horizons, and forms of multimodal evidence, allowing us to examine the behavior of multimodal memory integration across heterogeneous task settings rather than within a single benchmark or memory scenario.

Among these benchmarks, 8 constitute our primary evaluation suite, with 2 benchmarks selected from each task domain. We further include two additional conversation-understanding benchmarks, MemLens(Ren et al., 2026) and MemEye(Guo et al., 2026), for the multimodal sample analysis in Section 5.1, where we compare full-benchmark results with multimodal-only subsets to examine how instance-level multimodal necessity affects the observed benefits of multimodal memory.

Table 4: Overview of the 10 benchmarks used in our experiments, spanning 4 representative multimodal memory domains.

Domain	Benchmark	Samples	Output	Evaluator	Metric
GI	WebVoyager	514	Trajectory	LLM Judge	Success Rate
GI	Mind2Web	501	Trajectory	LLM Judge	Success Rate
VR	AgentVista	209	Trajectory	LLM Judge	Success Rate
VR	MMSearch-Plus	311	Trajectory	LLM Judge	Accuracy
CU	Mem-Gallery	1,711	Open-ended	LLM Judge	Accuracy
CU	MMRC	1,964	Open-ended	LLM Judge	Accuracy
CU	MemLens	789	Open-ended	LLM Judge	Accuracy
CU	MemEye	371	Open-ended	LLM Judge	Accuracy
VU	Video-MME	612	Multiple-choice	Exact Match	Accuracy
VU	HippoVlog	960	Multiple-choice	Exact Match	Accuracy

GI: GUI Interaction; VR: Visual Reasoning; CU: Conversation Understanding; VU: Video Understanding.

B.1 GUI Interaction

WebVoyager (He et al., 2024) is a benchmark for evaluating autonomous web agents on real-world websites. It consists of open-ended user instructions that require agents to interact with live webpages through multi-step navigation and action execution. The benchmark contains 643 tasks across 15 popular websites and evaluates whether an agent successfully completes the requested task using an automatic evaluator. We report task success rate as the evaluation metric. Since WebVoyager operates on live websites, a subset of the original tasks becomes unavailable or no longer reliably executable over time due to website updates, page removals, access restrictions, and changes in site structure. We therefore retain only tasks that remain accessible and executable under our evaluation environment. We partition tasks within each domain using a fixed random seed with one half to construct the memory while the other half to evaluate the agent's performance, resulting in 253 tasks and 261 tasks respectively, covering 12 websites. This partition method is applied to all of the GUI interaction and visual reasoning benchmarks.

Mind2Web (Deng et al., 2023) is a large-scale benchmark for generalist web agents that follow natural-language instructions to complete complex tasks on real-world websites. It contains over 2,000 open-ended tasks collected from 137 websites spanning 31 domains, together with crowd-sourced action sequences. The benchmark provides three generalization settings: cross-task, cross-website, and cross-domain. Since its original scale is substantially larger than the other benchmarks in our evaluation suite, we sample a subset of websites while preserving these evaluation settings to maintain a comparable overall evaluation budget. We report task success rate as the evaluation metric. In our experiments, we use 246 tasks for memory construction and 255 held-out tasks for evaluation, covering 22 websites.

B.2 Visual Reasoning

AgentVista (Su et al., 2026) evaluates multimodal agents on realistic visual tasks that require long-horizon reasoning and tool use. The benchmark contains 209 tasks grounded in 308 real-world images, spanning 7 major categories and 25 sub-domains. These tasks involve visually rich inputs such as photographs, screenshots, maps, and technical diagrams, and require agents to perform long-horizon reasoning with interleaved tool use, including visual grounding, external information retrieval, and intermediate verification. We evaluate agent performance using task success rate. We use the full benchmark in our experiments, with 102 tasks for memory construction and 107 held-out tasks for evaluation.

MMSearch-Plus (Tao et al., 2025) is a benchmark for multimodal browsing agents that emphasizes fine-grained visual reasoning during information search. It contains 311 tasks spanning 8 diverse categories, including geography, sports, academia, film/TV, technology, games, vlogs,

and music. Each task is designed to require the extraction and propagation of fine-grained visual cues through iterative image–text retrieval and cross-modal verification, making it difficult to solve through text-only search heuristics. We evaluate the final task outcome using accuracy. We use the full benchmark in our experiments, with 154 tasks for memory construction and 157 held-out tasks for evaluation.

B.3 Conversation Understanding

We consider 4 benchmarks for long-term multimodal conversation understanding. Mem-Gallery and MMRC constitute the primary evaluation benchmarks, while MemLens and MemEye are additionally included in the multimodal sample analysis in Section 5.1 to compare full-benchmark performance with multimodal-only subsets.

Mem-Gallery (Bei et al., 2026) evaluates multimodal long-term conversational memory over coherent multi-session dialogues with tightly coupled textual and visual information. Information is progressively introduced, referenced, and updated across sessions, requiring agents to maintain and reason over evolving multimodal memory. The benchmark covers memory extraction and adaptation, memory reasoning, and memory knowledge management, with 1,711 evaluation QA pairs constructed from 240 conversation sessions in 20 scenes. We evaluate agent performance using LLM-judge accuracy. Full datasets are used in our experiments.

MMRC (Xue et al., 2025) is a large-scale benchmark for evaluating multimodal memory and reasoning in sustained real-world conversations. It contains 5,120 carefully selected conversations and 28,720 manually annotated questions, covering six core abilities: information extraction, cross-turn reasoning, information update, image management, memory recall, and answer refusal. Compared with template-based conversational benchmarks, MMRC is collected from open-ended real-world interactions and contains substantially longer and more diverse multimodal conversations. We evaluate agent performance using LLM-judge accuracy. Since only 1964 QA pairs from 425 conversation sessions are publicly available, we use this subset in our experiments.

MemLens (Ren et al., 2026) is a benchmark for evaluating multimodal long-term memory in multi-session conversations. It contains 789 questions covering five core memory abilities: information extraction, multi-session reasoning, temporal reasoning, knowledge update, and answer refusal. The benchmark is instantiated at four standardized context lengths ranging from 32K to 256K tokens, enabling controlled evaluation under different long-context settings. A key feature of MemLens is that many questions are explicitly constructed to require joint reasoning over visual and textual evidence rather than admitting a text-only shortcut. We evaluate agent performance using LLM-judge accuracy. Full datasets of the 32K version are used in our experiments.

MemEye (Guo et al., 2026) is a visual-centric benchmark for evaluating multimodal long-term memory in image-grounded conversations. It organizes memory difficulty along two complementary dimensions: the granularity of decisive visual evidence, ranging from scene-level to pixel-level details, and the reasoning depth required to use the retrieved evidence, ranging from single-evidence retrieval to relational association and evolutionary synthesis. The benchmark contains 371 mirrored multiple-choice and open-ended questions across 8 life-scenario tasks, with explicit validation for visual necessity and shortcut resistance. We evaluate agent performance using LLM-judge accuracy. Full datasets are used in our experiments.

B.4 Video Understanding

Video-MME (Fu et al., 2025) is a comprehensive benchmark for evaluating multimodal models on video understanding across diverse temporal scales and visual domains. It contains 900 videos totaling 254 hours, together with 2,700 manually annotated question–answer pairs. The benchmark covers 6 major visual domains and 30 subfields, with video durations ranging from 11 seconds to 1 hour, and further incorporates subtitles and audio in addition to visual frames. This design enables evaluation of both short- and long-range multimodal reasoning over heterogeneous video content. We evaluate agent performance with multiple-choice accuracy. To maintain a tractable and comparable evaluation budget, we use the short-video subset of Video-MME. We further exclude

videos without subtitles to ensure consistent textual information availability across all evaluated samples. The resulting subset contains 204 videos and 612 questions.

HippoVlog (Lin et al., 2026) is a benchmark for long-form audiovisual understanding with a particular focus on cross-modal associative memory. It is constructed from 25 daily vlogs totaling 682 minutes and contains 1,000 manually validated questions across four memory-function categories: cross-modal binding, auditory-focused retrieval, visual-focused retrieval, and semantic reasoning. Unlike conventional video benchmarks that primarily test sequential comprehension, HippoVlog emphasizes reconstructing relevant multimodal experiences from partial cues and reasoning over evidence distributed across extended audiovisual streams. We evaluate agent performance with multiple-choice accuracy. We filter out one video which was no longer available and the resulting subset contains 24 videos and 960 questions.

C Implementation Details

C.1 Model Parameters

For Qwen3-VL-32B-Instruct, GPT-5.4-nano, and Gemini-2.5-Flash, we use the default decoding configurations provided by their respective serving endpoints. For Qwen3.5-27B, we set the temperature to 0.7, top-p to 0.8, and top-k to 20, with thinking disabled following the official recommendation. We use `qwen3-vl-embedding` with an embedding dimension of 1024 for all retrieval settings, and `gpt-5.4-mini` as the LLM judge throughout all experiments. All models are accessed through their official APIs.

C.2 Memory Implementations

ReasoningBank. We adapt ReasoningBank (Ouyang et al., 2026) to WebVoyager, Mind2Web, AgentVista, and MMSearch-Plus. Each task is allowed up to 15 environment interaction steps. After a trajectory is completed, the corresponding benchmark evaluator determines the task outcome, after which a success- or failure-specific prompt distills up to three non-overlapping and generalizable memory items. Each memory item follows a fixed *Title–Description–Content* schema, where the description contains one sentence and the content contains one to three sentences. The extraction prompt discourages task-specific website names, queries, and literal string values, encouraging reusable strategies rather than verbatim storage of individual trajectories.

All four memory levels are constructed from the same source trajectories, while differing only in the multimodal information made available at the designated memory stages.

At **L1**, memory construction is entirely text-based. The extractor only receives the task instruction, actions, textual observations, and task outcome, while visual observations are excluded. Memory retrieval is performed using textual query–memory similarity, and the retrieved memories are injected into the agent context in textual form.

At **L2**, the memory extractor additionally receives the last k screenshots of the trajectory ($k = 5$ by default), while the resulting memory representation, retrieval process, and memory injection remain text-only. We use a fixed recency window of five screenshots to capture visual states closest to task completion or failure, such as the final interface state, visible feedback, and other outcome-relevant UI cues. Restricting the visual input to this fixed window also limits redundant visual context and maintains a consistent construction budget across benchmarks.

At **L3**, multimodal information is further introduced into memory storage and retrieval. Each retrieval query consists of the current task instruction and its initial visual state, while each memory key consists of the historical task instruction paired with an initial trajectory screenshot. Queries and memory keys are encoded into fused text–image representations and ranked by cosine similarity. We retrieve the top-3 memories by default. The retrieved memory values themselves remain textual and are provided to the agent without raw images during downstream reasoning.

At **L4**, the retrieved visual evidence is additionally exposed to the agent during memory utilization. Along with the textual memory contents, we attach the corresponding raw memory images to the agent context, with at most two images provided for each retrieved memory item.

Omni-SimpleMem. We implement Omni-SimpleMem for the four conversation-understanding benchmarks. Each dialogue turn is stored as an individual memory record, while Mem-Gallery additionally stores one profile record when a persona summary is available. To prevent caption leakage in Mem-Gallery, we exclude the original image captions while retaining the original dialogue text and raw images. Each memory record maintains a concise summary for memory injection, a retrieval-oriented text field, and a richer evidence field stored in a cold store. Near-duplicate records are archived through similarity-based selective ingestion.

We use a hybrid retrieval pipeline that combines dense similarity, BM25 lexical matching, metadata-based matching, and lightweight entity-link expansion. The base retrieval budget is $k = 3$ for Mem-Gallery and $k = 8$ for MMRC; the retriever may expand this budget for question types that require evidence from multiple turns, such as counting or list questions. Retrieved summaries are further expanded with richer cold-store evidence subject to a fixed context budget.

All four memory levels are built from the same underlying dialogue histories, differing only in the modalities available at the designated memory stages.

At **L1**, each dialogue turn is represented solely from its textual content. The summary, retrieval text, and detailed evidence are all constructed from the dialogue text, and retrieval uses text-only query and memory representations. The downstream model receives only retrieved textual evidence.

At **L2**, the memory constructor additionally receives images associated with each dialogue turn. A multimodal prompt jointly processes the dialogue text and images to generate a concise summary, a detailed evidence description, and a retrieval-oriented textual representation. Thus, visual information can enrich the stored textual memory, while retrieval and downstream memory injection remain text-only.

At **L3**, multimodal information is further introduced into memory storage and retrieval. Each memory key is represented by the retrieval-oriented text paired with the images associated with its source dialogue turn. When the current question contains an image, its text and image are jointly encoded as the retrieval query; otherwise, the textual question alone is used. Multimodal dense similarity is incorporated into the same hybrid retrieval pipeline used at lower levels, together with BM25 lexical matching, metadata-based matching, and lightweight entity-link expansion. Candidate memories are then re-ranked using the combined retrieval signals. The retrieved memory records remain textual when injected into the downstream model.

At **L4**, we retain the same multimodal construction and retrieval process as L3, while additionally exposing the raw images associated with the retrieved dialogue turns to the downstream model.

WorldMM. We implement WorldMM for the two video-understanding benchmarks. The original WorldMM framework maintains episodic, semantic, and visual memories over multiple temporal granularities. In our experiments, we retain the episodic and visual memory components while disabling semantic-triple memory, so that the comparison isolates the effect of multimodal information without introducing an additional structured-memory mechanism.

Episodic memory is constructed at multiple temporal scales, including 3s, 10s, 30s, 1min, and 10min, with the active scales selected according to video duration. Each episodic memory entry contains a timestamped textual description of the corresponding video segment. Visual memory stores temporally localized video clips together with pre-computed visual embeddings. For episodic retrieval, candidates are first retrieved independently at each temporal scale and then globally ranked, after which the top three entries are retained. For visual memory, we retrieve the top-3 relevant clips.

We use progressive retrieval with at most 3 reasoning rounds. At each round, the agent formulates a refined search query conditioned on the current question and previously retrieved evidence, and terminates retrieval once sufficient evidence has been collected.

All four memory levels operate on the same source videos and differ only in the multimodal information available during memory construction, storage and retrieval, and final answer generation.

At **L1**, episodic memory is constructed solely from the original video subtitles. Both memory indexing and retrieval rely on textual representations, and the agent receives only timestamped textual evidence during reasoning.

At **L2**, visual information is additionally incorporated during memory construction through precomputed video captions and multi-scale summaries. The resulting episodic memory therefore contains textual descriptions of visual events, objects, attributes, and actions. Retrieval and final answer generation remain text-only.

At **L3**, multimodal information is further introduced into memory storage and retrieval through visual memory. Episodic and visually retrieved candidates are jointly considered under the same final evidence budget as L2, with at most $k = 3$ evidence units exposed to downstream reasoning. Instead of exposing raw frames, L3 provides the textual descriptions associated with the retrieved clips, including timestamps, captions, subtitles, and summaries, to the agent for reasoning.

At **L4**, we retain the same multimodal construction and visual retrieval process as L3, while additionally providing raw visual evidence from the retrieved clips to the final answer model. Although multiple candidate frames may be available for each retrieved clip, the final reasoning context is capped at two raw images to keep the visual input budget controlled and avoid overwhelming the agent with excessive visual information.

C.3 Repeated Runs and Failure Handling

To reduce variance introduced by stochastic model behavior and long-horizon environment interaction, we repeat each experimental configuration 3 times and report the average performance across the three valid runs. This is particularly important for GUI interaction and visual reasoning tasks, where multi-step execution can introduce additional variability.

Executions affected by infrastructure failures unrelated to the agent's decision-making, such as network interruptions, API request failures, or temporarily unavailable services, are treated as invalid runs and are rerun. Such failures are excluded only when they prevent the intended experimental procedure from being completed; ordinary task failures produced by the agent are retained as part of the evaluation.

D Additional Experimental Results

D.1 Robustness across MLLM Backbones

To examine whether our findings generalize beyond Qwen3-VL-32B-Instruct and GPT-5.4-nano, we repeat the L1–L4 evaluation on 2 additional models: Qwen3.5-27B, and Gemini-2.5-Flash. These backbones span different model families, parameter scales, and deployment settings, covering both open-source and proprietary MLLMs. Figures 5–6 present the complete results for the remaining 2 backbones across all 8 primary benchmarks. Together with the Qwen3-VL-32B-Instruct and GPT-5.4-nano results reported in the main paper, this yields a complete evaluation over 4 MLLM backbones, 8 benchmarks, and 4 multimodal memory levels.

As shown in Figures 5–6, the overall pattern remains consistent with the main results: increasing multimodal integration depth does not lead to uniformly better performance. Although absolute scores vary substantially across backbones, deeper integration does not consistently outperform shallower configurations. The exact best-performing level can also shift with the underlying model, indicating that backbone capability affects the magnitude of the gain or loss introduced by multimodal memory. Nevertheless, none of the evaluated MLLMs exhibits a universal preference for deeper multimodal integration across all benchmarks.

To further examine whether the effect of each integration stage is stable across backbones, we compare the direction of performance change for three adjacent transitions: L1→L2, L2→L3, and L3→L4. Table 5 summarizes whether each transition consistently improves performance, consistently degrades performance, or exhibits model-dependent behavior across the four evaluated backbones.

Despite variation in the exact optimal level, several adjacent transitions show stable effects across different backbones. For example, the transition from L3 to L4 consistently hurts performance on WebVoyager, Mind2Web, AgentVista, and Mem-Gallery, whereas the same transition consistently benefits Video-MME. Similarly, introducing multimodal information during memory construction (L1→L2) consistently benefits Video-MME and HippoVlog, while consistently degrading perfor-

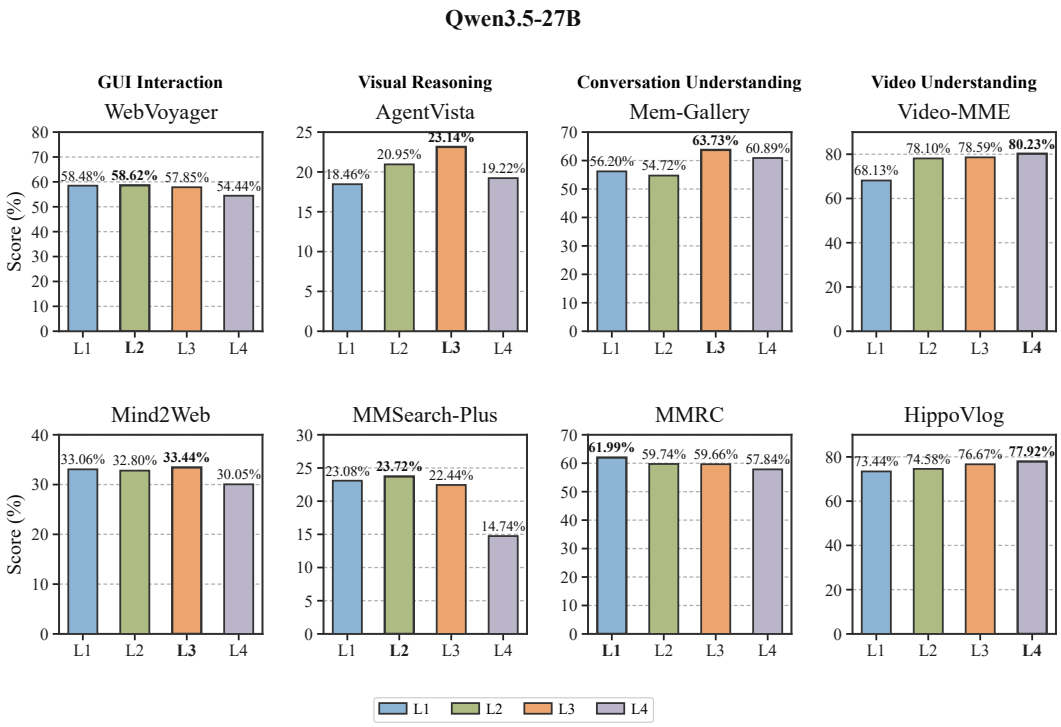


Figure 5: Performance across 4 multimodal memory integration levels on 8 benchmarks using Qwen3.5-27B. Each column corresponds to one task domain, with 2 benchmarks per domain. Bold values indicate the best-performing memory level for each benchmark.

Table 5: Cross-backbone consistency of adjacent multimodal integration transitions. ↑ and ↓ denote consistent improvement and degradation across all four backbones, respectively, while *Mixed* indicates model-dependent effects.

Benchmark	**L1→L2**	**L2→L3**	**L3→L4**
WebVoyager	Mixed	Mixed	↓
Mind2Web	Mixed	Mixed	↓
AgentVista	Mixed	Mixed	↓
MMSearch-Plus	↑	Mixed	↓
Mem-Gallery	Mixed	↑	↓
MMRC	↓	Mixed	↓
Video-MME	↑	Mixed	↑
HippoVlog	↑	Mixed	Mixed

mance on MMRC. These patterns suggest that the choice of backbone primarily affects the magnitude of the gain or loss, whereas the direction of certain stage-specific effects is more strongly associated with the memory requirements of the benchmark.

Overall, the backbone study reveals two complementary forms of robustness. First, the absence of a monotonic relationship between multimodal integration depth and performance persists across all evaluated MLLMs. Second, several adjacent integration transitions exhibit consistent positive or negative effects across backbones, indicating that these effects are not artifacts of a particular model. At the same time, other transitions remain model-dependent, showing that the exact optimal integration depth can still interact with backbone capability. Together, these results further support the view that multimodal integration should be analyzed at the level of individual memory stages rather than treated as a single scalar notion of "more multimodality."

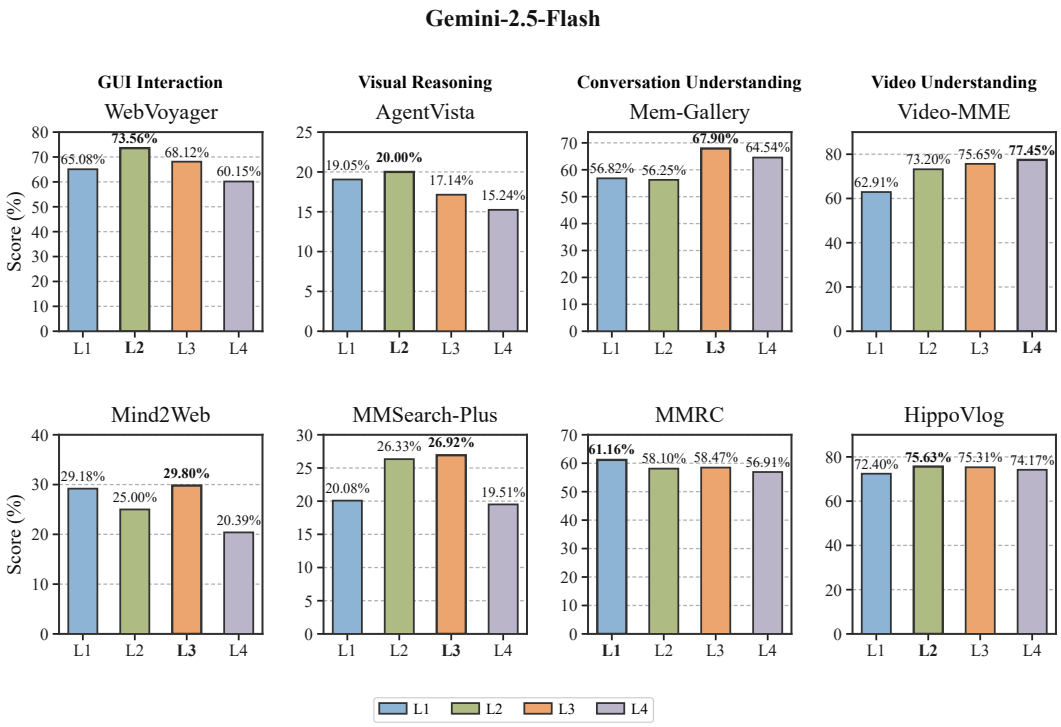


Figure 6: Performance across 4 multimodal memory integration levels on 8 benchmarks using Gemini-2.5-Flash. Each column corresponds to one task domain, with 2 benchmarks per domain. Bold values indicate the best-performing memory level for each benchmark.

D.2 Results of Alternative Memory Implementations

The main paper evaluates alternative implementations of the same multimodal memory levels using Qwen3-VL-32B-Instruct, showing that the concrete mechanism used to realize a given integration level can affect performance. To examine whether this implementation sensitivity is specific to a particular backbone, we repeat the same implementation-replacement study with Qwen3.5-27B on four representative benchmarks: WebVoyager, AgentVista, Mem-Gallery, and Video-MME.

For each comparison, we replace only the implementation of the corresponding memory level while keeping the underlying memory framework, source data, backbone, and other experimental configurations unchanged. The alternative implementation preserves the same multimodal integration depth as the original one, but realizes the corresponding memory stage using a different concrete mechanism. This allows us to separate the effect of *where* multimodal information is introduced in the memory pipeline from *how* that stage is implemented.

As shown in Figure 7, changing the concrete implementation within the same memory level can lead to non-negligible performance differences. The effect is particularly visible on AgentVista: replacing the L3 implementation decreases performance from 23.14% to 19.47%, whereas replacing the L4 implementation increases performance from 19.22% to 22.35%. This confirms that a multimodal integration level does not uniquely determine performance; the specific mechanism used to instantiate that level also matters.

Importantly, however, the alternative implementations do not systematically outperform the original ones. Their effects vary across both memory levels and benchmarks. On WebVoyager, the alternative L3 and L4 implementations improve performance, while the alternative L2 implementation performs slightly worse. On AgentVista, the alternative L4 implementation improves performance, whereas the alternative L3 implementation substantially degrades it. Similarly, the alternative L4 implementation slightly improves Mem-Gallery but decreases performance on Video-MME. Thus, replacing the implementation changes the magnitude and sometimes the direction of the performance difference, but does not yield a universally superior alternative.

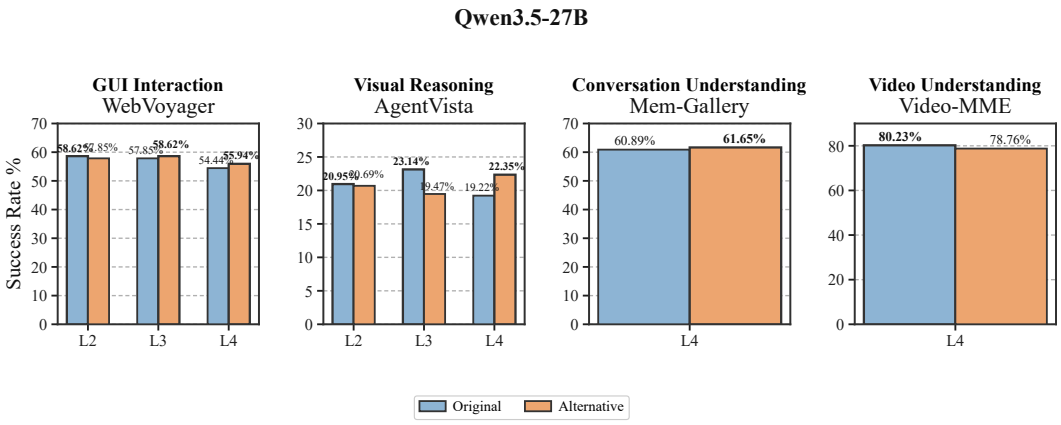


Figure 7: Sensitivity to alternative memory implementations on Qwen3.5-27B. For each memory level, *Alternative* replaces the concrete implementation while preserving the corresponding multimodal integration depth. Bold values indicate the better implementation within each pair.

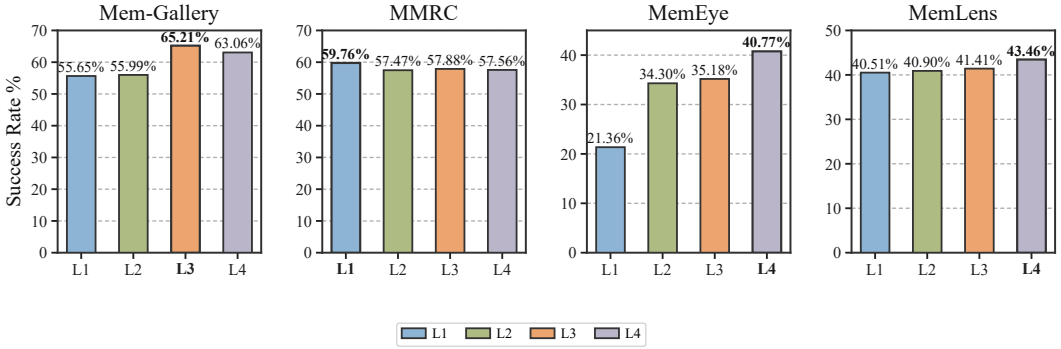


Figure 8: Performance across 4 multimodal memory integration levels on the full versions of four conversation-understanding benchmarks using Qwen3-VL-32B-Instruct. Bold values indicate the best-performing memory level for each benchmark.

Together with the Qwen3-VL-32B-Instruct results reported in the main paper, these experiments show that implementation choice is an important source of performance variation across multiple backbones. At the same time, changing the implementation does not make deeper multimodal integration consistently preferable or produce a single implementation that dominates across tasks. These results therefore suggest that multimodal integration depth and concrete memory implementation should be treated as two distinct design dimensions: the former determines *where* multimodal capability is introduced in the memory pipeline, while the latter determines *how* that capability is realized.

D.3 Detailed Results on Multimodal Sample Composition

Section 5.1 shows that aggregate benchmark evaluation can obscure the benefit of multimodal memory when text-only instances are mixed with genuinely multimodal ones. Here, we provide the complete L1–L4 results for four conversation-understanding benchmarks, comparing performance on the full benchmarks with their multimodal-only subsets.

The full-benchmark results provide the reference point for analyzing how sample composition affects the observed benefit of multimodal memory.

Controlling for multimodal sample composition. Not every instance in a multimodal benchmark necessarily requires multimodal information. When text-only instances are mixed with genuinely multimodal ones, aggregate performance can understate the benefit of multimodal memory. We

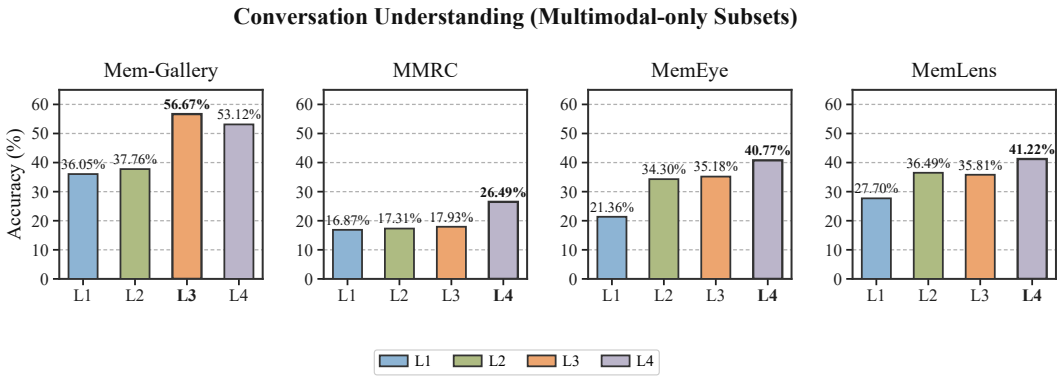


Figure 9: Performance across four multimodal memory integration levels on the multimodal-only subsets of four conversation-understanding benchmarks with Qwen3-VL-32B-Instruct. Bold values indicate the best-performing memory level for each benchmark.

therefore exclude non-multimodal samples and evaluate only the multimodal subsets of the four benchmarks. Specifically, we retain the TTL, VR, and VS subsets of Mem-Gallery (817 questions), the IM subset of MMRC (403 questions), all 371 multimodal samples from MemEye, and 37 multimodal samples from MemLens.

Figure 9 shows that restricting evaluation to multimodal-only samples changes the observed benefits of multimodal memory. In particular, MMRC shifts from preferring L1 on the full benchmark to preferring L4 on the multimodal subset. More generally, all four multimodal-only subsets benefit from introducing multimodal information beyond the text-only L1 configuration. Comparing the full-benchmark and multimodal-only results therefore shows that sample composition can substantially affect the observed benefit of multimodal memory. When many instances do not require multimodal evidence, gains on genuinely multimodal samples can be diluted in aggregate evaluation.

Overall, these results show that aggregate benchmark evaluation can underestimate the benefit of multimodal memory when text-only instances are mixed with genuinely multimodal ones. Evaluating multimodal-only subsets provides a clearer view of when multimodal memory is actually useful.

E Incremental Relevance of Retrieved Historical Experience

Section 5.2 argues that, in transfer-oriented memory settings, the usefulness of memory is constrained by whether retrieved past experience provides information that is actually transferable to the current task. We therefore conduct a retrieval audit on WebVoyager, Mind2Web, AgentVista, and MMSearch-Plus to quantify the *incremental relevance* of the historical experiences retrieved by the memory system.

Our goal is deliberately different from evaluating retrieval similarity or memory quality. Given a current task together with the information already available to the agent, we ask whether a retrieved historical trajectory contains additional task-specific information that can reduce uncertainty, constrain the reasoning or action space, or otherwise facilitate solving the current task. We evaluate the underlying historical trajectory rather than its distilled memory representation, thereby separating the *applicability of the retrieved experience itself* from the quality of a particular memory construction or representation mechanism.

Audit units. For each benchmark, we collect the memories actually retrieved by the L1 system during evaluation. Let

$$x_i = (q_i, c_i)$$

denote a target task, where q_i is the task goal or question and c_i contains the task information already available before historical memory is introduced. For every retrieved memory item m_{ij}, we use its source metadata to recover the historical trajectory h_{ij} from which the memory was constructed.

Score	Level	Criterion
0	None	The history provides no useful information beyond the current-task context.
1	Generic	The history shares only broad domain, topic, tool, or procedural similarity, without materially reducing uncertainty for the target task.
2	Actionable	The history provides a concrete and non-obvious task-relevant clue that can meaningfully narrow the target task's reasoning or action space.
3	Directly reusable	The history provides strongly target-specific information that resolves an important uncertainty or can be directly reused for the target task.

Each audit unit is therefore a target–history pair

$$a_{ij} = (x_i, h_{ij}),$$

representing one historical experience that was actually retrieved for target task i.

Current-task and historical context. For WebVoyager and Mind2Web, the current-task context includes the task goal, website information, and initial textual environment state; for AgentVista and MMSearch-Plus, it includes the current question and input images. The historical context contains the source task goal or question and its recorded trajectory, including textual actions and observations. We use the original trajectory rather than the distilled memory item to isolate the applicability of the historical experience itself from the quality of memory construction.

Conditional incremental relevance. We use `gpt-5.4-mini` with temperature 0 as the judge. For each audit unit $a_{ij} = (x_i, h_{ij})$, the judge assesses the information contributed by h_{ij} *conditional on* the current-task context x_i already being available. We denote this judgment by

$$\rho(x_i, h_{ij}) \in \{0, 1, 2, 3\}.$$

The score is intended to capture *incremental transfer value*, rather than surface similarity between the two tasks. In particular, sharing the same website, topic, tool, or generic interaction pattern is not sufficient for a high score.

The key distinction is between scores 1 and 2: generic similarity is not sufficient for actionable relevance. A score of 2 or above requires information that provides concrete guidance beyond what is already available in the current task context.

Aggregate measures. Let

$$\mathcal{A} = \{(x_i, h_{ij})\}$$

denote the set of valid audited target–history pairs. We summarize the audit using the two quantities reported in Table 1.

The MEAN RELEVANCE is the empirical expectation of the conditional relevance score:

$$\mathrm{MR} = \frac{1}{|\mathcal{A}|} \sum_{(x,h)\in\mathcal{A}} \rho(x, h).$$

The ACTIONABLE RELEVANCE is the proportion of retrieved histories that cross the actionable threshold:

$$\mathrm{AR} = \frac{1}{|\mathcal{A}|} \sum_{(x,h)\in\mathcal{A}} \mathbf{1}[\rho(x, h) \geq 2]\,.$$

MEAN RELEVANCE captures the overall amount of incremental task-specific information contained in retrieved history, whereas ACTIONABLE RELEVANCE measures how often retrieval returns an experience that can materially constrain the current task rather than merely share superficial or generic similarity.

This audit therefore isolates a prerequisite for transfer-oriented memory: before richer multimodal representations can be useful, the retrieved historical experience must first be applicable to the current task. Low values indicate that the retrieved experience itself provides little useful guidance for the current task, regardless of how richly it is represented multimodally.

F PAIRED ANSWER-SUPPORT AUDIT

To quantify how adjacent memory levels change the answer evidence available to the agent, we conduct a paired answer-support audit on the evidence-oriented benchmarks analyzed in Section 5.3. All paired audit metrics are computed independently and averaged across 3 runs. We omit the run index below for notational simplicity. The audit complements the end-to-end task evaluation by asking a more local question: *does the additional multimodal capability introduced by the higher memory level make the reference answer better supported by the retrieved evidence?*

Paired evidence construction. For each query i, let q_i denote the question and y_i its reference answer. We construct a pair of evidence conditions

$$(E_i^-, E_i^+),$$

where E_i^- contains the evidence available under the lower memory level and E_i^+ contains the corresponding evidence under the adjacent higher level. The question, answer options when applicable, and reference answer are held fixed across the pair.

The exact comparison follows the capability introduced by each transition. For L1→L2, we compare the L1 textual memory with the vision-enriched textual memory produced at L2. For L2→L3, we compare the evidence actually retrieved under the two retrieval configurations. For L3→L4, we compare textual evidence with the same retrieved content augmented by the raw visual evidence exposed at L4. For L2→L3 and L3→L4, the evidence conditions are constructed to match the information actually available to the agent at the corresponding memory level. In Video-MME, the L3 condition consists of the episodic text available at L2 together with the additional text associated with visually retrieved clips. For L3→L4, the higher-level condition further augments the L3 textual evidence with raw visual evidence. Video-MME and HippoVlog use the frames from the retrieved clips, while Mem-Gallery and MMRC use the retrieved raw images within their runtime image budgets.

All evidence pairs are presented to the judge in randomized A/B order. The mapping from A/B to the lower and higher memory levels is hidden during judgment and restored only during post-hoc aggregation.

Answer-support judgment. We use `gpt-5.4-mini` with temperature 0 as the judge. Given a query, its reference answer, and one displayed evidence condition, the judge evaluates how strongly the displayed evidence alone supports the reference answer. It is explicitly instructed not to rely on outside knowledge, the original agent response, or the identity of the memory level.

We define the answer-support score

$$S_i = s(q_i, y_i, E_i) \in \{0, 1, 2, 3\},$$

using the following rubric:

Score	Level	Criterion
0	None	The evidence contains no information relevant to establishing the reference answer.
1	Related	The evidence is topically related to the question but provides no concrete clue toward the reference answer.
2	Partial	The evidence provides a useful answer-relevant clue, but remains insufficient to reliably determine or verify the reference answer.
3	Sufficient	The displayed evidence is sufficient to determine or strongly verify the reference answer.

The key distinction is between scores 2 and 3. A score of 3 indicates that the memory condition contains sufficient evidence for answering the query, whereas scores below 3 indicate that some answer-critical information is still missing.

For each query, we obtain

$$S_i^- = s(q_i, y_i, E_i^-), \qquad S_i^+ = s(q_i, y_i, E_i^+),$$

for the lower- and higher-level conditions, respectively. The judge additionally identifies the displayed text or visual evidence supporting its decision. Structured outputs are validated for score ranges and evidence references.

Evidence metrics. Let $\mathcal{A}$ denote the paired audit set. We first report the average answer-support scores under the lower- and higher-level conditions:

$$\textsc{Answer Support}^{-} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} S_i^{-}, \qquad \textsc{Answer Support}^{+} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} S_i^{+}.$$

These scores summarize the overall amount of answer-relevant evidence available before and after the memory transition, including partial changes that do not cross the sufficiency threshold.

We define Baseline Sufficiency as

$$\textsc{Baseline Sufficiency} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \mathbf{1}[S_i^{-} = 3],$$

which measures the fraction of queries whose lower-level evidence is already sufficient to determine or strongly verify the reference answer.

We further define Evidence Rescue and Evidence Loss as

$$\textsc{Evidence Rescue} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \mathbf{1}[S_i^{-} < 3 \wedge S_i^{+} = 3],$$

and

$$\textsc{Evidence Loss} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \mathbf{1}[S_i^{-} = 3 \wedge S_i^{+} < 3].$$

Evidence Rescue captures queries for which the higher memory level turns previously insufficient evidence into sufficient evidence, whereas Evidence Loss captures the reverse transition. Both quantities are normalized by the full paired audit set, and therefore reflect the prevalence of rescue and loss events rather than conditional rescue or loss rates.

Together, these metrics distinguish three complementary aspects of an adjacent memory transition: the overall amount of answer support, the extent to which evidence gaps are resolved, and the extent to which previously sufficient evidence is lost.

Relation to task performance. All of the evidence metrics above are computed on the complete benchmark evaluation sets detailed in Table 4.

Let $c_i^{-}, c_i^{+} \in \{0, 1\}$ denote correctness at the lower and higher memory levels, and let $\mathcal{D}$ denote the full benchmark evaluation set. We report

$$\textsc{Baseline Acc.} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} c_i^{-}, \qquad \Delta\textsc{Acc.} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} (c_i^{+} - c_i^{-}).$$

The paired audit and end-to-end evaluation therefore measure different but complementary quantities: the former characterizes how the available evidence changes across memory levels, while the latter measures whether those evidence changes ultimately translate into improved task performance.

G A Four-Level Taxonomic View of Existing Multimodal Memory Systems

We categorize mainstream agent memory frameworks according to where multimodal information is introduced in the memory pipeline, including memory construction, storage/retrieval, and utilization, following the 4-level hierarchy defined in Section 3.1.

Table 6 shows that existing memory systems occupy different positions along the multimodal integration hierarchy, ranging from text-only memory to fully multimodal utilization. This suggests that multimodal memory design is naturally stage-dependent rather than a single binary choice.

Table 6: Representative agent memory systems organized by our four-level multimodal integration hierarchy.

Domain	**L1** Text-only	**L2** Construction	**L3** Storage & Retrieval	**L4** Utilization
VR	–	XSkill MemVerse	ViLoMem MuseAgent	–
GI	ReasoningBank AWM	M2-Memory	MMSkill	MementoGUI
CU	–	Tele-Mem	M2A ChronoMem	Omni- SimpleMem
VU	–	VideoLucy NS-Mem MemVerse	M3-Agent	WorldMM

GI: GUI Interaction; VR: Visual Reasoning; CU: Conversation Understanding; VU: Video Understanding.

However, prior systems also differ in their memory representations, retrieval mechanisms, and task settings, making it difficult to isolate the effect of integration depth from cross-system comparisons. Our framework addresses this by progressively introducing multimodal capability within a controlled memory pipeline.

H LIMITATIONS

Our study instantiates the proposed multimodal memory hierarchy using representative memory frameworks for each task domain. While this controlled setup enables systematic comparison across multimodal integration levels, a single representative implementation cannot capture the full diversity of memory architectures and design choices within each domain. Although we additionally examine alternative implementations for selected memory levels, a broader evaluation across more memory frameworks would further strengthen the generality of our findings.

In addition, our analysis primarily focuses on the effect of multimodal integration on task performance and the functional role of multimodal information in the memory pipeline. Other practical considerations, such as inference cost, latency, memory footprint, and computational efficiency, are not systematically evaluated in this work. Investigating these systems-level trade-offs together with performance constitutes an important direction for future work.