HIVE-MEM: INTERLEAVING HORIZONTAL MEMORY EXPANSION AND VERTICAL EVIDENCE COMPOSITION FOR MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

As multimodal agents accumulate increasingly long interaction histories, effectively retrieving and utilizing past information becomes critical for long-term interaction. Existing memory systems typically rely on semantic retrieval over compressed memories and expose predefined representations, which may miss relevant memories and overlook heterogeneous evidence needs across queries. We introduce **HIVE-MEM**, a multimodal memory framework that interleaves **H**or**I**zontal Memory Expansion with **VE**rtical Evidence Composition. For horizontal memory expansion, HIVE-MEM organizes memory episodes into a relational graph using textual and visual anchors, expanding semantically retrieved seeds with related neighbors to recover complementary information. For vertical evidence composition, each retrieved memory retains multiple evidence forms, including compact summaries, captions, grounded visual views, raw images, and interaction history. A lightweight trainable selector independently composes an evidence subset for each retrieved memory, enabling query-adaptive utilization of heterogeneous multimodal information. This two-dimensional design jointly broadens memory coverage and adapts the evidence contributed by individual memories. Experiments across three multimodal memory benchmarks show that HIVE-MEM consistently improves task performance across base models while reducing token consumption by about 90% compared with the most efficient competing multimodal baselines.

1 INTRODUCTION

As multimodal agents increasingly engage in persistent interactions involving both language and visual observations, memory becomes essential for preserving and reusing information from past interactions over long horizons (Hu et al., 2025; Huang et al., 2026). Existing memory systems typically transform interaction histories into compact representations and retrieve relevant entries when needed, allowing agents to access past information beyond the immediate context (Kang et al., 2025; Xu et al., 2025; Fang et al., 2025; Sun et al., 2023). In multimodal settings, however, useful information can span multiple modalities and levels of granularity (Bei et al., 2026; Liu et al., 2026a), making it difficult for a fixed memory representation and retrieval process to consistently surface the information most useful for future queries (Liu et al., 2026b; Wang & Chen, 2025; Long et al., 2026). This raises a fundamental challenge: *how can multimodal memory systems flexibly identify and utilize relevant information from heterogeneous interaction histories at inference time?*

Despite recent progress, existing multimodal memory systems still face two key limitations. *First, structured memory is often centered on abstract entities and relations rather than interaction episodes themselves*. While such structures are effective for consolidating knowledge, they may fail to capture cross-episode contextual affinities in multimodal histories, where related episodes can share partial cues across text and images, such as entities, scenes, or events, without forming explicit symbolic relations (Liu et al., 2025; 2026b). As a result, retrieval may still miss relevant episodes beyond directly matched memories, especially when useful information is distributed across multiple interactions. *Second, retrieved memories are typically exposed through predefined representations*, despite their heterogeneous evidence requirements. While compact textual abstractions may suffice for some memories, others require fine-grained visual or contextual evidence, making a fixed

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.

multimodal representation either insufficient or unnecessarily detailed, thereby limiting memory utilization and increasing token consumption (Long et al., 2026; Wang & Chen, 2025). These limitations call for a memory framework that can both recover relevant information distributed throughout multimodal histories and adaptively compose the evidence exposed by each retrieved memory.

To address these challenges, we propose **HIVE-MEM**, which broadens semantic retrieval through *Horizontal Memory Expansion* and tailors the multimodal representation of each retrieved memory through *Vertical Evidence Composition*. Horizontally, HIVE-MEM organizes memory episodes into a lightweight relational graph using textual and visual anchors extracted from multimodal histories. Starting from semantically retrieved seeds, it expands retrieval toward associated episodes through cross-episode contextual affinities, complementing direct semantic matching with relational signals. Vertically, HIVE-MEM retains multiple evidence forms for each memory, including compact summaries, captions, grounded visual views, raw images, and interaction history. A lightweight RL-trained selector independently composes the evidence for each retrieved memory conditioned on the query, allowing different memories to expose different evidence subsets while jointly considering answer quality and token consumption. Together, these two dimensions broaden relevant recall while adapting multimodal evidence to the needs of individual memories. Experiments across three multimodal memory benchmarks and two base models show consistent improvements over existing multimodal methods with substantially lower token consumption. Further analyses validate the complementary benefits of horizontal expansion and adaptive evidence composition, demonstrate flexible behavior across query and budget settings, and show only marginal runtime overhead.

Our main contributions are summarized as follows:

- We propose **HIVE-MEM**, a multimodal memory framework that jointly improves memory retrieval and evidence utilization, enabling flexible use of long text–image interaction histories beyond fixed retrieval and representation pipelines.
- We realize this framework via complementary mechanisms. *Horizontal Memory Expansion* augments semantic retrieval with cross-episode contextual affinities, while *Vertical Evidence Composition* adaptively composes multimodal evidence per retrieved memory.
- Extensive experiments across three multimodal memory benchmarks and two base models demonstrate improvements over multimodal methods with substantially lower token consumption. Further analyses validate both components and flexible evidence adaptation.

2 RELATED WORK

2.1 TEXT-CENTRIC AGENT MEMORY

LLM agents increasingly rely on external memory to preserve and reuse information beyond bounded context windows (Zhong et al., 2024; Packer et al., 2023). A major line of work transforms interaction histories into compact and reusable representations through extraction, compression, and consolidation (Chhikara et al., 2025; Xu et al., 2025; Kang et al., 2025; Fang et al., 2025; Liu et al., 2026c). Beyond independent memory entries, structured approaches organize past experiences through entities, relations, or associative links, allowing retrieval to exploit dependencies among memories rather than relying solely on isolated similarity matching (Zhu et al., 2026a). Another line introduces learned memory management, where policies or evolving operations adapt how memories are constructed, updated, and reused from downstream feedback (Yan et al., 2025; Zhang et al., 2026a; Yu et al., 2026; Wang et al., 2025; Zhang et al., 2026b). Together, these methods have substantially advanced long-term memory for text-based agents, but remain primarily designed around textual information, leaving heterogeneous multimodal evidence less explored.

2.2 MULTIMODAL AGENT MEMORY

Recent work extends agent memory beyond text to persistent visual and multimodal interaction histories. MIRIX (Wang & Chen, 2025) maintains specialized memory types for heterogeneous user information, while M3-Agent (Long et al., 2026) organizes episodic and semantic experience around entity-centric multimodal representations. M2A (Feng et al., 2026) complements raw interaction records with higher-level semantic memories, whereas MemVerse (Liu et al., 2025) consolidates multimodal experience into hierarchical knowledge structures. Recent systems further

explore contextual memory organization, reliability-aware retrieval, and unified multimodal memory pipelines (Jain et al., 2025; Lu et al., 2026; Liu et al., 2026b). Collectively, these methods establish effective mechanisms for storing, structuring, and accessing multimodal experience. In contrast, HIVE-MEM focuses on improving how constructed memories are utilized at inference time, by broadening retrieval through relations among memory episodes and adaptively composing the evidence exposed by each retrieved memory, enabling more effective multimodal memory use.

3 METHODOLOGY

Our methodology consists of three stages as shown in Figure 1. We first organize the interaction history into episode-centric memory representations with complementary textual and visual information in Sec. 3.1. We then introduce *Horizontal Memory Expansion* in Sec. 3.2 to complement semantic retrieval with cross-episode contextual associations. Finally, Sec. 3.3 presents *Vertical Evidence Composition*, which adaptively determines the evidence exposed by each retrieved memory.

3.1 EPISODE-CENTRIC MULTIMODAL MEMORY FORMATION

We consider a growing interaction history $\mathcal{H} = \{x_t\}_{t=1}^{T}$, where each interaction may contain textual content and associated images. Rather than treating every turn independently, we partition the history into contiguous *interaction episodes* (formed over fixed-length token spans), offering a practical granularity while preserving local context. Each episode e_i is subsequently transformed into a *memory episode* m_i, which serves as the basic memory object in HIVE-MEM.

To instantiate each memory episode, we derive a compact summary s_i with vision-language models (VLMs), providing a concise semantic representation for subsequent retrieval. To preserve salient information beyond this coarse abstraction, we associate each episode with two sets of fine-grained, VLM-generated modality-aware anchors that capture complementary cues from language and vision. Specifically, the *textual anchors* $\mathcal{T}_i$ capture salient semantics from the interaction text, such as entities, events, and topics, while the *visual anchors* $\mathcal{V}_i$ record visually grounded cues, such as objects, scenes, activities, and attributes, from the associated images. Visual anchors are useful for retaining visually salient information that may be only partially captured by textual abstractions. Together, the two anchor sets provide lightweight modality-specific metadata beyond the summary and are used to characterize contextual affinities between memory episodes in Sec. 3.2.

Beyond these compact descriptors, each memory episode retains multiple evidence forms at different levels of detail. By default, these include the raw interaction text r_i, compact summary s_i, image captions c_i, raw images o_i, and a grounded visual view g_i, which we introduce to localize context-relevant regions in the original images while suppressing irrelevant and noisy background, inspired by (Liu et al., 2026d). Accordingly, each memory episode is represented as

$$m_i = \big(r_i,\, s_i,\, c_i,\, o_i,\, g_i,\, \mathcal{T}_i,\, \mathcal{V}_i\big). \tag{1}$$

This episode-centric representation retains complementary views of multimodal interaction history rather than reducing each memory to a single abstraction. The compact summary supports semantic retrieval, the modality-aware anchors provide cues for modeling cross-episode affinities, and the evidence forms preserve complementary textual and visual details. Together, these components support the *Horizontal Memory Expansion* and *Vertical Evidence Composition* introduced next.

3.2 HORIZONTAL MEMORY EXPANSION

Semantic retrieval is effective at identifying memories directly aligned with the current query, but it largely considers each memory in isolation. In multimodal histories, relevant context may also be connected through weaker cues distributed across text and images, while explicit entity–relation structures primarily capture symbolic relations. To complement these signals, we model *cross-episode contextual affinities* between memory episodes. A graph provides a natural substrate for preserving such soft associations without requiring predefined relation types, allowing retrieved memories to serve as entry points for discovering additional relevant context.

Cross-Episode Affinity Graph. We construct an undirected affinity graph $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, where each node corresponds to a memory episode m_i. We characterize pairwise affinity through overlaps

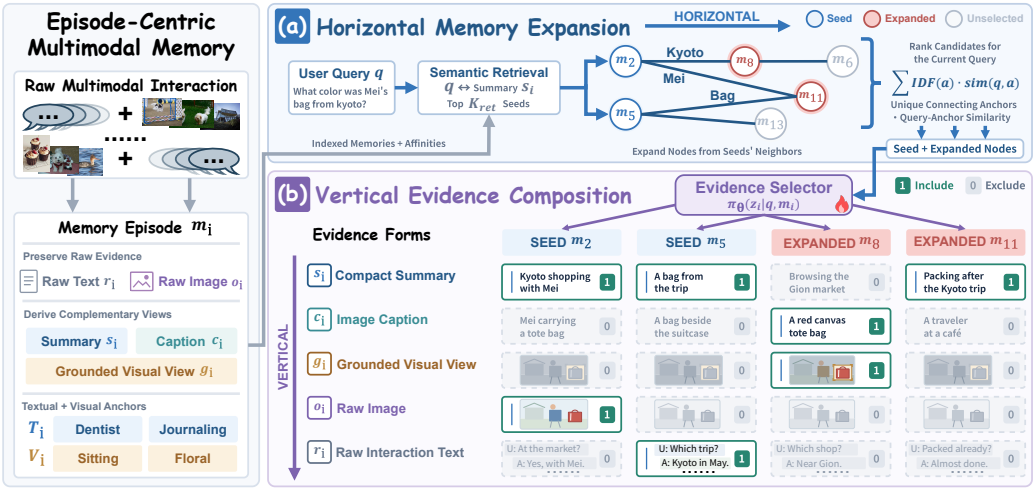

Figure 1: Overview of HIVE-MEM. **(a) Horizontal Memory Expansion** retrieves seed episodes through summary-based semantic matching and expands to graph neighbors ranked by the specificity and query relevance of their connecting anchors. **(b) Vertical Evidence Composition** uses an RL-trained selector to make binary inclusion decisions over compact summaries, image captions, grounded visual views, raw images, and raw interaction text, composing a query-dependent evidence subset for each retrieved episode to balance answer quality and token consumption.

among the textual and visual anchors defined in Sec. 3.1. These anchors preserve fine-grained semantic cues beyond summary-level similarity, particularly cross-modal correspondences where a concept is expressed textually in one episode but appears visually in another. For two memory episodes m_i and m_j, we define their matching anchor set as

$$\mathcal{A}_{ij} = (\mathcal{T}_i \cap \mathcal{T}_j) \cup (\mathcal{T}_i \cap \mathcal{V}_j) \cup (\mathcal{V}_i \cap \mathcal{T}_j). \tag{2}$$

The first term captures shared textual semantics, while the latter two form cross-modal bridges between language and vision. To reflect the varying specificity of matched anchors, we weight them by inverse document frequency (IDF), which suppresses frequent cues while emphasizing more distinctive ones. For each $a \in \mathcal{A}_{ij}$, with N memory episodes, we define

$$\mathrm{df}(a) = \sum_{k=1}^{N} \mathbb{I}[a \in \mathcal{T}_k \cup \mathcal{V}_k], \qquad \mathrm{IDF}(a) = \log \frac{N+1}{1+\mathrm{df}(a)}, \qquad w_{ij} = \sum_{a \in \mathcal{A}_{ij}} \mathrm{IDF}(a). \tag{3}$$

The resulting w_{ij} emphasizes more distinctive shared anchors while downweighting ubiquitous ones. To keep the graph sparse, we retain at most K_{edge} edges per node by greedily selecting candidate edges in descending w_{ij} order while respecting the degree limit.

Affinity-Guided Memory Expansion. Given a query q, we first retrieve K_{ret} seed memories by measuring semantic similarity between q and each compact summary s_i, and collect their one-hop neighbors, excluding already retrieved memories, as expansion candidates. To prioritize neighbors based on query-relevant anchor overlap, we score the fine-grained anchors they share with the seed memories rather than reuse summary-level similarity. For a candidate m_u, let $\mathcal{B}_u$ denote the unique anchors shared with any seed memory. Its expansion score is defined as

$$S_{\text{exp}}(m_u \mid q) = \sum_{a \in \mathcal{B}_u} \mathrm{IDF}(a)\ \mathrm{sim}(q,a). \tag{4}$$

where $\mathrm{sim}(q,a)$ denotes the cosine similarity between the query and anchor embeddings. The IDF term captures anchor specificity, while the similarity term measures query relevance. We rank expansion candidates by S_{exp} and select the top K_{exp}, yielding up to $K_{\text{ret}} + K_{\text{exp}}$ retrieved memory episodes that combine direct semantic relevance with query-relevant cross-episode affinities.

3.3 VERTICAL EVIDENCE COMPOSITION

After horizontal expansion determines which memory episodes to retrieve, the remaining question is how each retrieved memory should be presented to the downstream model. A single evidence form is unlikely to suit all memories in multimodal settings. Compact summaries and captions are efficient but may omit fine-grained details, while raw interaction text and images preserve richer information at higher inference cost. Moreover, the appropriate evidence can vary across both queries and retrieved memories. We therefore formulate memory utilization as *per-memory evidence composition*, where each retrieved memory contributes a query-dependent subset of its available evidence.

Query-Adaptive Evidence Composition. For each retrieved memory m_i, we consider candidate evidence forms $\mathcal{E}_i = \{r_i, s_i, c_i, o_i, g_i\}$ by default, corresponding to raw interaction text, compact summary, image caption, raw image, and grounded visual view (note that this formulation is not limited to these forms and can accommodate additional evidence types). Rather than selecting a single representation, we introduce a lightweight trainable selector to jointly determine which evidence forms should be exposed. For each memory, the selector takes the concatenated embeddings of the query and compact summary, $\mathbf{h}_i = [E(q); E(s_i)]$, as its lightweight input. A single forward pass jointly predicts evidence-specific selection probabilities

$$\mathbf{p}_i = f_\theta(\mathbf{h}_i) = (p_{i,1}, \ldots, p_{i,|\mathcal{E}_i|}) \in [0,1]^{|\mathcal{E}_i|}, \tag{5}$$

where $p_{i,k}$ denotes the selection probability of the k-th evidence form. These probabilities parameterize binary decisions, collected as $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,|\mathcal{E}_i|})$, where $z_{i,k} \in \{0,1\}$ indicates whether the corresponding evidence form is included. Applying the selector to each retrieved memory allows different episodes to expose different evidence subsets for the same query.

Learning Evidence Composition. Determining the optimal evidence subset is non-trivial because its utility is ultimately reflected by downstream answer quality, while the selection itself is discrete. We therefore optimize the selector with reinforcement learning (Schulman et al., 2017) using a reward that jointly considers answer quality and evidence consumption. Specifically, for the composed retrieval context and resulting answer $\hat{y}$, we define

$$R = R_{\text{ans}}(\hat{y}, y) - \lambda \widetilde{C}(\mathbf{z}), \qquad C(\mathbf{z}) = \sum_i \sum_k z_{i,k}\, \tau_{i,k}, \qquad \widetilde{C}(\mathbf{z}) = \text{Norm}(C(\mathbf{z}))\,. \tag{6}$$

where R_{ans} measures answer quality, $C(\mathbf{z})$ denotes token consumption of selected evidence, $\widetilde{C}(\mathbf{z})$ its normalized value, and λ controls the balance between answer quality and inference cost. Optimizing answer quality alone favors richer evidence, whereas increasing λ encourages more selective compositions under tighter computation budgets. This allows HIVE-MEM to adapt evidence usage across different resource constraints without changing the underlying memory representation.

Together, both dimensions of HIVE-MEM operate complementarily. *Horizontal Memory Expansion* determines which memory episodes to recall, while *Vertical Evidence Composition* determines what information each retrieved memory contributes. Their combination enables query-adaptive use of multimodal interaction history in both memory coverage and evidence granularity.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmarks and Metrics. We evaluate HIVE-MEM on three multimodal memory benchmarks: Mem-Gallery (Bei et al., 2026), WorldMemArena (Liu et al., 2026a), and H2HMem (Zhu et al., 2026b), covering long-horizon interaction histories with both textual and visual information. For task performance, we report token-level F1 and an LLM-as-a-judge score (LJ), where GPT-4o-mini is used as the evaluator. To measure inference efficiency, we additionally report total token consumption (Tok.), summing input and output tokens. Further details on evaluation protocols and judge prompts are provided in Appendix A.3.

Baselines. We compare HIVE-MEM against two groups of representative memory methods. For *(1) text-centric memory*, we include **LightMem** (Fang et al., 2025), **SimpleMem** (Liu et al., 2026c), and **BudgetMem** (Zhang et al., 2026b), which cover lightweight memory construction, compression, runtime adaptation, and learned memory management. For *(2) multimodal memory*, we compare

Table 1: **Main results on multimodal memory benchmarks.** We report F1 and LLM-Judge (LJ) for task performance, together with total token consumption (Tok.). Avg. denotes the average across three benchmarks. Best and second-best results among multimodal methods are shown in **bold** and underlined; text-centric baselines are included for reference only, as they operate under a different modality setting without image processing during memory-bank construction.

	Mem-Gallery			WorldMemArena			H2HMem			Avg.		
Method	**F1**	**LJ**	**Tok.**	**F1**	**LJ**	**Tok.**	**F1**	**LJ**	**Tok.**	**F1**	**LJ**	**Tok.**
Qwen3-VL-4B-Instruct												
Text-centric memory												
LightMem	27.70	38.55	0.547	21.24	45.00	0.394	16.33	26.40	0.357	21.75	36.65	0.433
SimpleMem	31.76	36.27	0.462	16.31	38.50	0.598	18.30	24.11	0.433	22.12	32.96	0.498
BudgetMem	34.89	42.45	0.264	22.01	56.82	0.361	20.32	29.49	0.272	25.74	42.92	0.299
Multimodal memory												
M2A	47.95	62.55	11.707	14.12	37.73	24.536	23.20	31.39	19.233	28.42	43.89	18.492
MIRIX	39.16	51.55	6.267	20.72	60.68	12.334	23.09	31.11	7.266	27.66	47.78	8.622
MMA	33.51	51.55	33.034	**28.24**	59.77	6.343	22.82	33.74	7.879	28.19	48.35	15.752
MemVerse	29.00	44.64	5.445	22.06	62.73	11.916	21.36	28.89	6.074	24.14	45.42	7.812
HIVE-MEM	**49.44**	**71.55**	**0.491**	23.66	**68.18**	**0.991**	**23.87**	**34.17**	**0.492**	**32.32**	**57.96**	**0.658**
GPT-5-mini												
Text-centric memory												
LightMem	43.08	51.45	0.264	22.32	57.73	0.419	24.40	29.51	0.363	29.93	46.23	0.348
SimpleMem	33.65	48.09	0.303	21.66	54.27	0.493	21.32	32.29	0.461	25.54	44.89	0.419
BudgetMem	41.11	48.09	0.224	23.61	61.59	0.411	22.36	28.82	0.355	29.02	46.17	0.330
Multimodal memory												
M2A	49.79	65.82	9.311	23.00	65.68	7.955	29.34	42.28	12.558	34.04	57.93	9.941
MIRIX	39.89	55.64	4.543	22.50	77.50	12.518	26.23	37.71	4.988	29.54	56.95	7.350
MMA	33.87	54.55	4.228	24.32	67.50	3.450	25.30	37.82	19.433	27.83	53.29	9.037
MemVerse	45.66	63.91	8.855	24.03	78.64	10.818	29.16	39.58	12.782	32.95	60.71	10.818
HIVE-MEM	**53.31**	**72.45**	**0.367**	**24.80**	**79.32**	**1.021**	**29.71**	**42.29**	**0.856**	**35.94**	**64.69**	**0.748**

Note: Tok. denotes total input and output token consumption, reported in millions (M).

against **M2A** (Feng et al., 2026), **MIRIX** (Wang & Chen, 2025), **MMA** (Lu et al., 2026), and **MemVerse** (Liu et al., 2025), which explicitly model and retrieve information from multimodal interaction histories.

Implementation Details. We initialize the evidence selector as a lightweight multi-layer perceptron (MLP) and use Proximal Policy Optimization (PPO) as the default reinforcement learning algorithm. Experiments use Qwen3-VL-4B-Instruct and GPT-5-mini as base models, with GPT-5-mini used only for transfer evaluation (i.e., the selector is only trained on Qwen3-VL-4B-Instruct). Interaction histories are segmented into episodes of 256 tokens, and Qwen3-VL-Embedding-2B is used for semantic retrieval. By default, HIVE-MEM retrieves $K_{\text{ret}} = 5$ seed memories, expands $K_{\text{exp}} = 2$ additional memories, and sets the maximum graph degree to $K_{\text{edge}} = 4$, while all baselines use a consistent Top-7 retrieval setting. Unless otherwise specified, we set $\lambda = 0$ and further analyze the effect of budget constraints in Section 4.3. We split each benchmark into training, validation, and test sets with a 6:2:2 ratio. Appendix A provides additional implementation and evaluation details.

4.2 COMPARISON EXPERIMENTAL RESULTS

HIVE-MEM consistently improves task performance across models and benchmarks. As shown in Table 1, text-centric methods generally trail multimodal baselines, highlighting the value of preserving visual information, whereas existing multimodal methods benefit from richer context but often rely on less selective memory utilization. Against this backdrop, HIVE-MEM achieves the best average F1/LJ scores of 32.32/57.96 and 35.94/64.69 with Qwen3-VL-4B-Instruct and GPT-5-mini, respectively, establishing consistent gains over multimodal baselines. The advantage is not driven by a particular benchmark, as HIVE-MEM obtains the best LJ across all three benchmarks with the former backbone and the best F1/LJ across all three with the latter. This improvement supports the benefit of jointly broadening relevant recall and adapting per-memory evidence rather than relying

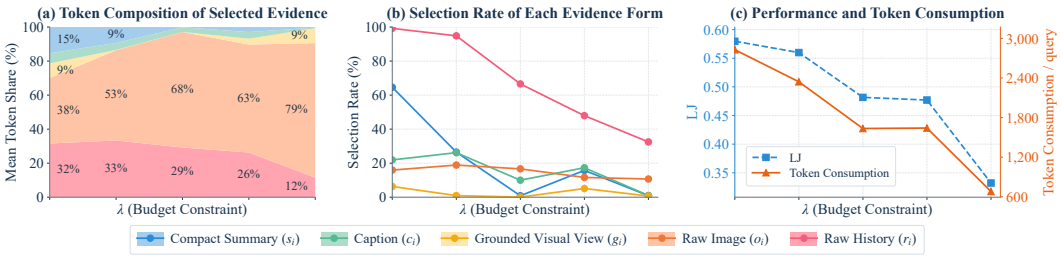

Figure 2: **Evidence composition under varying token-consumption preferences.** Results are averaged across Mem-Gallery, WorldMemArena, and H2HMem using Qwen3-VL-4B-Instruct. **(a)** Token shares of selected evidence forms, normalized to 100%. **(b)** Selection rates of individual evidence forms. **(c)** LJ and token consumption per query under different λ values.

on fixed retrieval and representation schemes. Notably, the latter results use the selector transferred directly from the former without retraining, further indicating cross-model generalization.

HIVE-MEM achieves these gains with substantially lower multimodal token consumption. Its average consumption is only 0.658M and 0.748M tokens under the two base models, compared with 7.812M and 7.350M for the most efficient competing multimodal baselines, corresponding to reductions of 91.6% and 89.8%. This efficiency gap partly arises because existing multimodal methods, despite organizing richer memories through multiple stores, hierarchical structures, or reliability-aware retrieval, still expose substantial retrieved context during answering. In contrast, HIVE-MEM composes evidence separately for each retrieved memory, retaining fine-grained textual or visual information only when useful. This enables substantially lower token consumption without sacrificing performance. For completeness, we also include text-centric methods, whose token usage reflects a different modality setting that omits image processing during memory-bank construction.exclude image processing during memory-bank construction.

Overall, the results show that HIVE-MEM improves multimodal memory performance while reducing inference context. Its consistent gains across benchmarks and backbones further support the benefits of jointly expanding memory coverage and adapting evidence composition.

4.3 How Does HiVe-Mem Adapt Evidence Composition?

To understand how the learned selector adapts under different budget constraints, we vary λ and analyze evidence usage with Qwen3-VL-4B-Instruct, averaging results across the three benchmarks. Figure 2 shows a clear shift toward more selective compositions as λ increases. The selection rate of raw history drops substantially, while several auxiliary evidence forms are progressively filtered out, indicating that the policy does not reduce cost uniformly across evidence types. Interestingly, raw images are selected relatively infrequently but contribute an increasingly large fraction of consumed tokens, reflecting their high per-selection cost and the tendency to retain them only when their visual detail is valuable. This contrast between selection frequency and token share highlights the importance of optimizing evidence utility jointly with consumption.

Figure 2(c) further shows that varying λ produces distinct quality–token-consumption operating points, allowing HIVE-MEM to explicitly control how aggressively evidence is selected under different budget constraints. As λ increases, the selector shifts toward more compact evidence compositions and lower token consumption, while smaller values retain richer evidence to prioritize answer quality. Overall, this budget-aware composition enables HIVE-MEM to flexibly adapt to different token budgets without changing the underlying memory system.

4.4 Ablation Study

We ablate the key designs of HIVE-MEM using Qwen3-VL-4B-Instruct and report LJ averaged equally across the three benchmarks in Table 2. For Horizontal Memory Expansion, *w/o Horizontal Expansion* replaces graph expansion with semantic retrieval while keeping the same number of retrieved memories, *w/o Visual Anchors* removes visual anchors from graph construction, and *w/o IDF Weighting* assigns shared anchors without IDF-based weighting. For Vertical Evidence Composi-

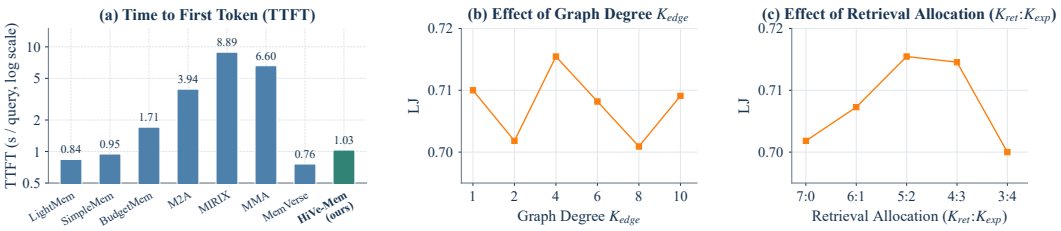

Figure 3: **Latency and hyperparameter sensitivity analysis** with Qwen3-VL-4B-Instruct on Mem-Gallery. **(a)** Time to first token (TTFT) across methods on a logarithmic scale. **(b)** Effect of graph degree K_{edge} on LJ with default $K_{\text{ret}} : K_{\text{exp}} = 5 : 2$. **(c)** Effect of retrieval allocation on LJ with default $K_{\text{ret}} + K_{\text{exp}} = 7$ and $K_{\text{edge}} = 4$.

tion, *w/o Vertical Composition* represents every retrieved memory using only its compact summary, while the remaining variants individually remove one evidence form from the composition space.

Both horizontal expansion and its relational signals contribute to retrieval quality. Disabling Horizontal Memory Expansion reduces LJ from 57.96 to 56.51, showing that retrieving additional memories through cross-episode affinities provides information beyond an equally sized semantic retrieval set. Removing visual anchors also degrades performance to 57.30, supporting the value of visual cues for connecting multimodal episodes. Notably, removing IDF weighting causes a larger drop to 56.55, indicating that distinguishing informative anchors from frequently occurring ones is important for constructing useful episode relations.

Adaptive evidence composition is particularly important for multimodal memory utilization. Replacing Vertical Evidence Composition with compact summaries alone causes the largest degradation, from 57.96 to 44.47, showing that summary-only representations are insufficient for many queries. Among individual evidence forms, removing raw text and raw images leads to the largest drops, reaching 49.68 and 51.02 LJ, respectively, highlighting the importance of preserving fine-grained information from both modalities. Compact summaries, captions, and grounded visual views also provide consistent gains when retained. Together, these results reveal that different evidence forms contribute complementary information across diverse multimodal memory queries, motivating query-adaptive composition rather than relying on compact summaries alone.

Table 2: **Ablation study of HiVe-Mem** averaged equally across Mem-Gallery, WorldMemArena, and H2HMEM using Qwen3-VL-4B-Instruct. The best LJ score is in **bold**.

Horizontal Memory Expansion		*Vertical Evidence Composition*			
Variant	**LJ**	**Variant**	**LJ**	**Variant**	**LJ**
HiVe-Mem (Full)	**57.96**	**HiVe-Mem** (Full)	**57.96**	**HiVe-Mem** (Full)	**57.96**
w/o Horizontal Expansion	56.51	w/o Vertical Composition	44.47	w/o Grounded Visual View	57.19
w/o Visual Anchors	57.30	w/o Compact Summary	56.86	w/o Raw Image	51.02
w/o IDF Weighting	56.55	w/o Image Caption	56.12	w/o Raw Text	49.68

4.5 Extensive Analysis

Runtime Efficiency Analysis. We evaluate time to first token (TTFT) on Mem-Gallery to examine the online overhead introduced by different memory systems. As shown in Fig. 3(a), several multimodal baselines incur substantially higher TTFT because answering involves additional runtime memory operations, such as agent coordination, memory decomposition, or evidence processing before generation. In contrast, HiVe-Mem introduces only lightweight graph expansion and evidence composition, resulting in a TTFT of 1.03 seconds. Although slightly higher than the simplest retrieval-based baselines, this overhead remains marginal and is substantially lower than other multimodal systems in practice. This suggests that the additional adaptivity of HiVe-Mem can be obtained without materially delaying response generation.

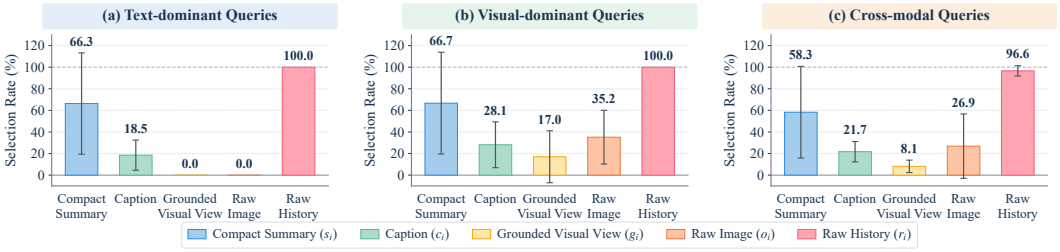

Figure 4: **Evidence selection across query types.** Using Qwen3-VL-4B-Instruct at $\lambda = 0$, we group the native query categories of each benchmark into **(a)** text-dominant, **(b)** visual-dominant, and **(c)** cross-modal queries, and report the selection rate of each evidence form. Rates are computed over all retrieved Top-7 memories and averaged equally across Mem-Gallery, WorldMemArena, and H2HMem. Error bars show ± 1 population standard deviation across benchmarks.

Sensitivity Analysis. Figures 3(b) and (c) examine the two key hyperparameters of Horizontal Memory Expansion. Performance remains stable across a wide range of K_{edge}, with the best result at the default $K_{\text{edge}} = 4$, indicating that HIVE-MEM is not sensitive to graph sparsity within a reasonable range. For retrieval allocation, keeping $K_{\text{ret}} + K_{\text{exp}} = 7$ reveals a balance between semantic retrieval and graph expansion. Starting from pure semantic retrieval, increasing K_{exp} consistently improves LJ and reaches its peak around 5:2, with similar performance at 4:3. Further increasing the expansion ratio eventually hurts performance, likely because too few semantic seeds provide a narrower and less diverse basis for discovering useful neighbors. These results show that reserving part of the retrieval budget for graph expansion improves over pure semantic retrieval, while retaining enough semantic seeds remains important for supplying diverse and reliable expansion candidates. Overall, HIVE-MEM remains robust across a broad range of retrieval configurations.

Evidence Selection across Query Types. To examine whether HIVE-MEM adapts evidence composition to heterogeneous information needs, we group the native query categories of each benchmark into text-dominant, visual-dominant, and cross-modal queries. For example, Visual-centric Reasoning (VR) and Visual-centric Search (VS) in Mem-Gallery are treated as visual-dominant, while Test-Time Learning (TTL) represents cross-modal queries. Similarly, H2HMem distinguishes text-oriented categories such as Temporal Reasoning from visual and cross-modal categories such as Reference & Evolution Tracking and Cross-modal Related Retrieval.

Figure 4 reveals clear query-dependent evidence patterns. Text-dominant queries almost never invoke grounded visual views or raw images, whereas visual-dominant and cross-modal queries select visual evidence substantially more often. At the same time, compact summaries remain frequently selected across all three query types, while raw history provides a stable textual backbone. Notably, even visual-dominant queries select raw images for only 35.2% of retrieved memories, indicating that visual information is not indiscriminately exposed whenever a query involves vision. Instead, compressed or grounded evidence can often provide sufficient information, with raw images retained only when additional visual detail is useful. Cross-modal queries exhibit an intermediate composition that combines textual and visual evidence, further demonstrating that HIVE-MEM adapts evidence at a finer granularity than simply routing queries by modality. Overall, these distinct selection profiles support the flexibility of per-memory evidence composition under heterogeneous query requirements.

5 CONCLUSION

We introduced HIVE-MEM, a multimodal memory framework that improves both memory retrieval and evidence utilization over long text–image interaction histories. *Horizontal Memory Expansion* complements semantic retrieval with cross-episode contextual affinities, while *Vertical Evidence Composition* adaptively selects heterogeneous evidence for each retrieved memory through reinforcement learning. Together, these two dimensions enable more flexible use of multimodal memory across different information needs and inference budgets. Experiments across multimodal memory benchmarks demonstrate consistent improvements over existing methods, with further analyses validating the effectiveness of both memory expansion and adaptive evidence composition.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and the identification of relevant literature. All AI-assisted text was reviewed and revised by the authors, and identified references were manually verified for correctness and relevance. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the memory construction, Horizontal Memory Expansion, and Vertical Evidence Composition in Secs. 3.1–3.3. Experimental settings, including benchmarks, baselines, model configurations, and evaluation metrics, are described in the main text. Additional details on memory episode construction, training and reward design, and evaluation protocols are provided in Appendices A.1, A.2, and A.3. We plan to release the source code and implementation upon publication to facilitate reproduction of our results.

REFERENCES

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.

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.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

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.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, et al. Rethinking memory mechanisms of foundation agents in the second half: A survey. *arXiv preprint arXiv:2602.06052*, 2026.

Jitesh Jain, Shubham Maheshwari, Ning Yu, Wen-mei Hwu, and Humphrey Shi. Augustus: An llm-driven multimodal agent system with contextualized user memory. *arXiv preprint arXiv:2510.15261*, 2025.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*, 2025.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, et al. Worldmemarena: Evaluating multimodal agent memory through action-world interaction. *arXiv preprint arXiv:2605.29341*, 2026a.

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*, 2026b.

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

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*, 2025.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026d.

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.

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

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: Towards llms as operating systems. 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-{\alpha}: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z Pan, Hinrich Schütze, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, 2026.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. Memskill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*, 2026a.

Haozhen Zhang, Haodong Yue, Tao Feng, Quanyu Long, Jianzhu Bao, Bowen Jin, Weizhi Zhang, Xiao Li, Jiaxuan You, Chengwei Qin, et al. Learning query-aware budget-tier routing for runtime agent memory. *arXiv preprint arXiv:2602.06025*, 2026b.

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.

Jinchang Zhu, Jindong Li, Cheng Zhang, Jiahong Liu, and Menglin Yang. Hela-mem: Hebbian learning and associative memory for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13757–13769, 2026a.

Shiping Zhu, Yibo Yang, Zhengyang Wang, Tiancheng Shen, Dandan Guo, and Ming-Hsuan Yang. H2hmem: A multimodal memory benchmark for agents in human-human interactions. *arXiv preprint arXiv:2606.09461*, 2026b.

APPENDIX CONTENTS

A Implementation Details **12**

A.1 Memory Episode 12

A.2 Training Details 13

A.3 Evaluation Details 13

A.4 Hyperparameters 14

B Extensive Experiments **14**

B.1 Evidence Composition under Transfer Evaluation 14

B.2 Evidence Selection across Query Types 14

C Limitations and Future Work **15**

D Broader Impact **15**

E Prompts **16**

E.1 Memory Construction Prompt 16

E.2 QA Prompts 17

E.3 LLM Judging Prompts 18

APPENDIX

A IMPLEMENTATION DETAILS

A.1 MEMORY EPISODE

Episode Construction. We construct memory episodes over approximately fixed-length token spans while respecting natural conversational boundaries. In particular, an episode is not forcibly truncated in the middle of a dialogue or session. If the nominal token boundary falls within an ongoing interaction segment, we extend the episode to the nearest valid boundary. This provides a consistent memory granularity while preserving local conversational continuity. Each raw episode is processed only once by a vision-language model (VLM), which jointly produces its compact summary together with the textual and visual anchors introduced in Sec. 3.1.

Grounded Visual View. To obtain a more focused visual representation, we construct the grounded visual view directly from the raw image and its associated interaction context. Specifically, the raw image and the relevant contextual passage are provided to a VLM, which identifies the target object discussed in context and predicts its spatial coordinates. We then crop the corresponding region from the original image to form the grounded visual view. Compared with the raw image, this localized view removes irrelevant background and concentrates visual content around the context-relevant object, yielding a more semantically focused representation. This design also reduces the visual input required by downstream models. **Depending on the benchmark, grounded visual views reduce token consumption by approximately 20–40% compared with the raw images.**

Finally, the evidence forms adopted in this work are not intended to define a fixed evidence space. HIVE-MEM treats evidence composition generically and can accommodate additional textual, visual, or derived representations. The raw interaction, compact summary, caption, raw image, and grounded visual view used here serve as representative evidence forms in our default instantiation.

A.2 Training Details

Cost Normalization. The task reward R_{ans} is directly computed from answer quality. For the cost term, we use the token consumption $C(\mathbf{z})$ of the selected evidence as the raw cost. To reduce sensitivity to its absolute scale, we maintain a sliding window of recent costs and normalize them using robust quantiles. Letting Q_5 and Q_{95} denote the 5th and 95th percentiles of $\sqrt{C}$ within the window, we compute

$$\widehat{C}(\mathbf{z}) = \text{clip}\left(\frac{\sqrt{C(\mathbf{z})} - Q_5}{Q_{95} - Q_5}, 0, 1\right). \tag{7}$$

Since answer-quality and cost signals may have different scales during training, we further align the normalized cost using

$$\alpha = \frac{\text{Std}(R_{\text{ans}})}{\text{Std}(\widehat{C}) + \epsilon}, \qquad \widetilde{C}(\mathbf{z}) = \alpha\, \widehat{C}(\mathbf{z}), \tag{8}$$

where the statistics are computed over recent training samples and ϵ is a small constant. The final reward is therefore

$$R = R_{\text{ans}} - \lambda \widetilde{C}(\mathbf{z}), \tag{9}$$

consistent with the formulation in Sec. 3.3.

PPO Optimization. For each query q, the per-memory evidence compositions $\mathbf{z}_i$ jointly form the complete evidence composition $\mathbf{z}$. During PPO training, the selector f_θ parameterizes a factorized policy over the binary evidence decisions through the evidence-specific probabilities predicted in a single forward pass. Specifically,

$$\pi_\theta(\mathbf{z}_i \mid \mathbf{h}_i) = \prod_k \pi_\theta(z_{i,k} \mid \mathbf{h}_i), \tag{10}$$

where each factor is parameterized by the corresponding selection probability $p_{i,k}$ defined in Sec. 3.3. The composed evidence is then used for downstream answering, yielding the reward R defined above.

Given the behavior policy $\pi_{\theta_{\text{old}}}$ used to collect the rollout, the likelihood ratio of the complete evidence composition is

$$\rho(\theta) = \prod_i \frac{\pi_\theta(\mathbf{z}_i \mid \mathbf{h}_i)}{\pi_{\theta_{\text{old}}}(\mathbf{z}_i \mid \mathbf{h}_i)}. \tag{11}$$

We estimate the advantage $\widehat{A}$ from the reward using a learned value baseline and maximize the standard clipped PPO objective

$$\mathcal{J}_{\text{PPO}} = \mathbb{E}\left[\min\left(\rho(\theta)\widehat{A}, \text{clip}(\rho(\theta), 1-\epsilon_{\text{ppo}}, 1+\epsilon_{\text{ppo}})\,\widehat{A}\right)\right], \tag{12}$$

where ϵ_{ppo} denotes the PPO clipping coefficient.

A.3 Evaluation Details

Benchmarks and Splits. We evaluate on three multimodal memory benchmarks, **Mem-Gallery** (Bei et al., 2026), **WorldMemArena** (Liu et al., 2026a), and **H2HMem** (Zhu et al., 2026b), which collectively cover long-horizon interaction histories involving both textual and visual information. Mem-Gallery evaluates memory over multimodal conversational histories, while WorldMemArena considers memory use over extended agent interactions. For WorldMemArena, we use only the *Lifelong* subset. H2HMem further evaluates memory over long multimodal dialogue histories. Across all benchmarks, we exclude adversarial settings and construct training, validation, and test sets using a 6:2:2 split.

Model Transfer Evaluation. The evidence selector is trained only with Qwen3-VL-4B-Instruct as the base model. For experiments with GPT-5-mini, we directly transfer the selector weights trained under Qwen3-VL-4B-Instruct and perform evaluation without additional training or adaptation. This setting allows us to examine whether the learned evidence-selection policy transfers across base models rather than relying on model-specific retraining.

For evaluation, we report token-level F1 and LLM-Judge (LJ), together with input and output token consumption. GPT-4o-mini is used as the judge model for LJ evaluation across all methods.

Table 3: Key hyperparameters used in our experiments.

Hyperparameter	Value	Description
Embedding and Retrieval		
Embedding model	Qwen3-VL-Embedding-2B	Multimodal encoder for queries and memory summaries.
Embedding dimension	2048	Dimension of embedding representations.
K_{ret}	5	Number of memories obtained by semantic retrieval.
K_{exp}	2	Number of memories added by horizontal expansion.
Evidence Selector		
Architecture	$2048 \to 256 \times 3$	Three-layer MLP with hidden dimension 256.
Selection threshold	0.5	Binary evidence-selection threshold at inference.
Initial action probability	0.5	Initial probability of selecting each evidence form.
PPO Training		
Learning rate	3×10^{-4}	Adam learning rate.
Clip ratio	0.2	PPO clipping coefficient.
Rollout / minibatch	$16/16$	Rollout and optimization minibatch sizes.
Update epochs	4	PPO updates per rollout batch.
Training epochs	6	Total training epochs.
γ / GAE λ	$1.0/1.0$	Discount factor and GAE coefficient.
Value / entropy coef.	$0.5/0.01$	Value-loss and entropy coefficients.
Max gradient norm	1.0	Gradient clipping threshold.
Reward		
Quality reward	F1	Answer-quality reward.
Inference		
Temperature	0.0	Deterministic generation.
Max output tokens	512	Maximum generation length per invocation.

A.4 Hyperparameters

We provide key hyperparameters used in our work in Table 3.

B Extensive Experiments

We further examine the behavior of HiVe-Mem under the GPT-5-mini transfer setting. Unless otherwise specified, the evidence selector trained with Qwen3-VL-4B-Instruct is directly transferred without retraining. These experiments complement the analyses in the main text by examining whether adaptive evidence composition remains effective across base models.

B.1 Evidence Composition under Transfer Evaluation

Figure 5 shows that the adaptive behavior observed with Qwen3-VL-4B-Instruct persists after transfer to GPT-5-mini. As λ increases, the selector does not uniformly reduce all evidence forms, but progressively changes the composition itself. Raw history is selected much less frequently under stronger constraints, while visual evidence remains selectively retained when useful. Although raw images are chosen for only a small fraction of retrieved memories, they account for a substantial share of consumed tokens whenever selected, again reflecting the heterogeneous cost of different evidence forms. Meanwhile, Fig. 5(c) exhibits distinct LJ–token-consumption operating points as λ varies. This demonstrates that the transferred selector preserves controllable evidence allocation and can adjust inference behavior to different token budgets without retraining.

B.2 Evidence Selection across Query Types

Figure 6 further reveals clear query-dependent evidence preferences. For text-dominant queries, grounded visual views and raw images are almost never selected, indicating that HiVe-Mem avoids unnecessary visual processing when textual evidence is sufficient. In contrast, visual-dominant and cross-modal queries invoke both grounded visual views and raw images substantially more often, with raw-image selection reaching 30.2% and 25.4%, respectively. Compact summaries remain frequently selected across all query types, while raw history serves as a consistently useful tex-

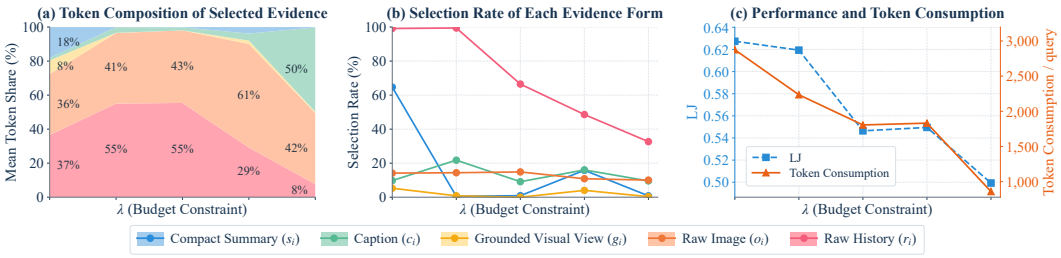

Figure 5: **Evidence composition under varying token-consumption preferences with GPT-5-mini.** Results are averaged across Mem-Gallery, WorldMemArena, and H2HMem using the transferred evidence selector. **(a)** Token shares of selected evidence forms, normalized to 100%. **(b)** Selection rates of individual evidence forms. **(c)** LJ and token consumption per query under increasing λ. Displayed settings are equally spaced in increasing λ order.

tual source. Importantly, visual evidence is still selected far below 100% even for visual-dominant queries, suggesting that relevant visual information is often already captured by compact or derived representations and that raw images are retained only when additional visual detail is needed. The consistent modality-dependent behavior under GPT-5-mini further supports the transferability and flexibility of per-memory evidence composition.

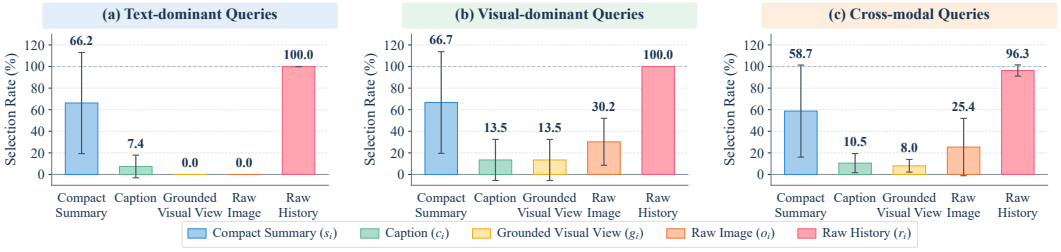

Figure 6: **Evidence selection across query types with GPT-5-mini.** Native query categories are grouped into **(a)** text-dominant, **(b)** visual-dominant, and **(c)** cross-modal queries. We report evidence-selection rates using the transferred selector at $\lambda = 0$. Rates are averaged equally across Mem-Gallery, WorldMemArena, and H2HMem, with error bars showing ± 1 population standard deviation across benchmarks.

C Limitations and Future Work

Although HiVe-Mem demonstrates consistent effectiveness across multimodal memory benchmarks, several extensions remain open. First, our experiments instantiate evidence composition with a representative set of textual and visual forms. Future work could incorporate richer derived evidence, such as structured visual descriptions or task-specific representations, within the same framework. Second, we adopt a lightweight MLP as the evidence selector for efficiency. Exploring alternative selector architectures may further improve evidence composition under different model and task settings. Finally, our evaluation considers two representative base models. Extending the study to a broader range of model families and scales would provide a more comprehensive understanding of the transferability of HiVe-Mem.

D Broader Impact

As multimodal agents evolve from short-lived question answering systems toward persistent interactive systems, the ability to selectively reuse long-term interaction history may become increasingly important. HiVe-Mem provides a step toward memory systems that do not treat stored multimodal information as a fixed representation, but instead adapt both what should be recalled and how retrieved information should be exposed for downstream reasoning. This perspective could support

a broad range of long-horizon applications, including personalized assistants, multimodal copilots, embodied agents, and interactive systems that continuously accumulate textual and visual context over time.

More broadly, adaptive memory utilization may become a fundamental capability for scalable agent systems. As interaction histories grow, processing all stored information uniformly becomes increasingly impractical. Mechanisms that dynamically recover relevant memories and compose only the necessary evidence could enable agents to maintain richer histories while controlling inference overhead. We hope HIVE-MEM encourages future research toward more flexible multimodal memory architectures that integrate additional modalities, richer memory representations, and increasingly adaptive mechanisms for organizing and reusing long-term interaction history.

E PROMPTS

E.1 MEMORY CONSTRUCTION PROMPT

Task Prompt Template

Role. You convert one conversation chunk and its attached images into exactly one structured memory record.
Input. `<conversation chunk>`, optionally `<attached original images>`, and optionally `<user name>`. The user name may only be used to resolve references to the user; no other profile facts may be used as summary or `Ti` evidence.
Instruction. Produce exactly one complete summary for the entire chunk, together with:

- `Ti`: attributes explicitly stated in the chunk text;
- `Vi`: attributes directly observable in the attached images.

The summary may combine relevant information from the text and images. Preserve names, numbers, dates, decisions, reasons, and image details needed for future question answering. If the chunk has a date, begin the summary with `On <session date>, .` Resolve relative dates when the chunk provides sufficient information.
Treat `Ti` and `Vi` as sparse retrieval properties. Include only concrete and distinctive attributes useful for linking the current chunk to other memories.
Constraint. Write exactly one non-empty summary and never split the chunk into multiple memory records. Do not add unsupported information.
A `Ti` attribute must be supported by words in the current chunk. Do not derive `Ti` from images or user-profile information. Do not convert questions, uncertain suggestions, conversation mechanics, or speaker metadata into asserted attributes.
A `Vi` attribute must be directly observable in an attached image. Do not derive `Vi` from the chunk text, image captions, or profile. Do not visually infer occupation, ownership, preference, causality, identity, date, or other non-observable facts. If no image is attached, return `"Vi":[]`.
Each attribute value must be an array of strings containing at most eight distinct values. Each of `Ti` and `Vi` may contain at most 16 attribute objects. Merge values that share the same entity and attribute, and do not repeat the same fact under multiple keys.
Allowed `Ti` attributes are: `relation`, `preference`, `occupation`, `trait`, `age`, `species`, `breed`, `owner`, `appearance`, `skill`, `status`, `category`, `color`, `style`, `material`, `use`, `kind`, `location`, `feature`, `role`, `date`, `participants`, `shape`, `texture`, `pattern`, `count`, `visible_state`, `action`, `pose`, `position`, `spatial_relation`, and `ocr_text`.
Allowed `Vi` attributes are: `color`, `appearance`, `shape`, `material`, `texture`, `pattern`, `count`, `visible_state`, `action`, `pose`, `position`, `spatial_relation`, and `ocr_text`.
If the same attribute is independently supported by both sources, include it once in `Ti` and once in `Vi`; never copy evidence from one source to the other.
Output. Return exactly one valid JSON object:

```
 {"summary":"<one complete summary of the entire
chunk>", "Ti":[{"entity":"<optional entity name>",
"attribute":"<allowed textual attribute>", "value":["<one or
more values>"]}], "Vi":[{"entity":"<optional visible entity>",
```

`"attribute":"<allowed visual attribute>", "value":["<one or more values>"]}]}`
Do not output a JSON array, multiple objects, multiple summaries, `MEMORY_ITEM` or `ENTITIES` lines, Markdown, or explanatory text.

E.2 QA PROMPTS

QA Instructions: Mem-Gallery

Answer from the provided multimodal conversation memory. Ground the answer in that memory, prefer the latest information when details change over time, and keep the response concise but complete. When the question asks for images, preserve the exact `image_id` values.

Additional user-message constraint for conflict detection:
The text inside `<answer>...</answer>` must be exactly "Yes." or "No."

Additional user-message constraint for visual search:
Return the exact matching `image_id` value or values. If several apply, sort them in ascending order and separate them with commas.

QA Instructions: WorldMemArena

Answer from the provided retrieved memories. Synthesize multiple entries when needed, use timestamps and the most recent evidence to resolve conflicts, and ground every claim in the memories. Convert relative dates using the memory timestamps. When images are requested, preserve exact `image_id` values. Keep the answer concise; if no relevant information is present, the answer text must be exactly "Not mentioned in memory."

QA Instructions: H2HMem

You are a memory testing system. `<question-type instruction>` Answer directly in English without reasoning, using no more than 100 words.

Additional system instruction for conflict detection:
The answer text must be strictly either Yes or No.

Additional system instruction for answer refusal:
If the information is absent, the answer text must be exactly "Not mentioned."

For H2HMem, the question-type instruction is selected from the following:

H2HMem Question-Type Instructions

Unimodal precise recall. Accurately recall the requested information from the conversation memory.
Cross-modal related retrieval. Use related textual and visual information from the conversation memory.
Knowledge resolution. Resolve knowledge consistently across the conversation memory.
Temporal reasoning. Reason about temporal relationships in the conversation memory.
Multimodal causal reasoning. Use textual and visual evidence for the requested causal reasoning.
Reference & evolution tracking. Track references and how they evolve across the conversation memory.
Test-time learning. Use knowledge introduced in the conversation memory to answer the question.
Conflict detection. Determine whether the information in the question conflicts with the conversation memory.
Answer refusal. Determine whether the question can be answered from the conversation memory.

E.3 LLM JUDGING PROMPTS

Mem-Gallery and H2HMem. These benchmarks share the following five-level rubric. In the output-format example, the reasoning placeholder is empty for Mem-Gallery and `<short explanation>` for H2HMem.

Judge Prompt: Mem-Gallery and H2HMem

You are an impartial judge evaluating the memory capabilities of an AI assistant with the question-answering task.
Your task is to compare the Assistant's Answer against the Ground Truth and assign a score of 0, 0.25, 0.5, 0.75, or 1.
Scoring Rubric.
Score 0 (Incorrect / Miss):

- The answer contradicts the Ground Truth.
- For Yes/No questions: The answer has the wrong polarity (e.g., says "Yes" when Ground Truth is "No").
- For Open-ended questions: The answer provides factually wrong information or hallucinations.
- The assistant fails to provide the required information.

Score 0.25 (Poor / Tangential):

- The answer touches on the topic but misses the **core entity** or key value required.
- The answer contains a mix of minor correct details and **significant hallucinations** or wrong associations.
- The answer is excessively vague to the point of being useless (e.g., answering "a dog" instead of "a golden retriever").

Score 0.5 (Partial / Vague):

- The answer is technically correct, but lacks confidence or is incomplete.
- The answer captures the **main entity or concept** correctly but misses a part of the required supporting details.
- For Yes/No questions: The polarity is correct, but the reasoning is flawed (if have), or the assistant is uncertain (e.g., "I think it might be Yes").
- For Open-ended questions: The answer is too general or misses key adjectives/details present in the Ground Truth.

Score 0.75 (Good / Minor Imperfection):

- The answer is largely accurate and captures the core information confidently.
- It misses only **minor details** (e.g., specific adjectives or secondary details) that do not alter the main truth.
- The answer contains all the correct information but includes unnecessary "fluff" or slight conversational filler that reduces precision.

Score 1 (Correct / Exact):

- The answer is accurate, precise, and confident.
- For Yes/No questions: The polarity matches the Ground Truth perfectly.
- For Open-ended questions: The answer contains **all** the core information and necessary details required by the Ground Truth without hallucinations.

Input Data.
Ground Truth: `<reference answer(s)>`
Assistant Answer: `<predicted answer>`
Output Format.
Output strictly in the following JSON format:

```
{"score": <0, 0.25, 0.5, 0.75, or 1>,
  "reasoning": "<reasoning placeholder>"}
```

WorldMemArena. The evaluator assigns a score of 1 to `Correct` and 0 to `Hallucination` or `Omission`. This evaluates answer correctness, not evidence coverage.

Judge Prompt: WorldMemArena

You are an **evaluation expert for AI memory system question answering**.
Based **only** on the provided **Reference Answer**, strictly evaluate the **accuracy** of the **Memory System Response**. Classify it as one of **Correct**, **Hallucination**, or **Omission**. Do **not** use any external knowledge or subjective inference.

Evaluation Criteria.

1. Correct

- The response accurately answers the question and is **semantically equivalent** to the Reference Answer.
- No contradictions with the Reference Answer.
- Synonyms, paraphrasing, and reasonable summarization are acceptable.

2. Hallucination

- The response includes information that **contradicts** the Reference Answer.
- When the Reference Answer is *unknown/uncertain*, yet the response provides a specific fact.

3. Omission

- The response is **incomplete** compared to the Reference Answer.
- It states "don't know" or "no related memory" even though the Reference Answer supplies the answer.
- For multi-element answers, missing **any** element counts as Omission.

Priority Rules.

- Both missing info AND fabricated info → **Hallucination**.
- No fabrication but missing info → **Omission**.
- Fully equivalent → **Correct**.

Information.
Reference Answer: `<reference answer(s)>`
Memory System Response: `<predicted answer>`

Output.

```
{
  "reasoning":  "Concise evaluation rationale",
  "evaluation_result":
    "Correct | Hallucination | Omission"
}
```