LEARNING TO ACCESS MULTIMODAL AGENT MEMORY: AGENTIC RETRIEVAL WITH STATE-LEVEL ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term multimodal memory enables agents to leverage temporally distributed evidence across interleaved conversations and images. Nevertheless, prevailing approaches largely rely on fixed retrieval strategies or rule-based agentic workflows, offering limited flexibility to adapt retrieval decisions as the available evidence evolves. Effective evidence acquisition requires evaluating evidence sufficiency and identifying unresolved information gaps before taking the next action. Accordingly, we propose state-level on-policy self-distillation (OPSD), in which privileged information (PI) explicitly captures these two factors. At each intermediate state of the student rollout, a verifier reconstructs these diagnoses by comparing the current short-term memory (STM) with the reference answer, providing privileged context for a frozen teacher to guide subsequent memory-access decisions. OPSD then aligns the student's action preferences with those of the privileged teacher, enabling evidence-dependent decisions on when to terminate and what evidence to acquire next without privileged information at inference time. To support this learning formulation, we develop a unified agentic memory-access framework that formulates evidence acquisition as sequential tool use, provides heterogeneous memory-access operations, and maintains a compact STM for consolidating intermediate evidence. We further construct MemVista, a training dataset comprising long-horizon multimodal conversations with temporally evolving facts and traceable supporting evidence. Experiments on Mem-Gallery demonstrate that our approach improves the mean LLM-judge score from 0.780 to 0.852 compared with the strongest external baseline. These results indicate that state-level OPSD with dynamic PI enables more effective memory access and substantially improves performance on long-term multimodal memory reasoning tasks.

1 INTRODUCTION

Long-term multimodal interactions produce extended histories of interleaved text and images. The accumulated conversations may exceed the capacity of a fixed context window, motivating the use of external memory systems (Packer et al., 2023; Hu et al., 2023; Zhong et al., 2024; Kang et al., 2025; Bei et al., 2026). However, accessing such memories remains challenging because relevant evidence may be distributed across multiple sessions, linked through temporal dependencies, or contained in visual details that are not fully captured by textual summaries (Bei et al., 2026). Effective long-term multimodal memory system therefore requires identifying and integrating complementary evidence across modalities and time, rather than simply retrieving isolated memory entries based on similarity.

One prominent line of research emphasizes memory organization and indexing, improving recall through the construction of persistent, searchable representations. Packer et al. (2023) and Kang et al. (2025) introduce hierarchical memory management, while Wang & Chen (2025) organize multimodal experiences into specialized memory components. Liu et al. (2026b) further combine semantic compression with multi-view indexing and query-aware retrieval, and Kang et al. (2026) introduce modality-dependent routing to improve access to multimodal evidence. Across these designs, semantic similarity matching remains an important mechanism for locating candidate memories. However, relevance does not imply evidence sufficiency: a retrieved memory item may be

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.

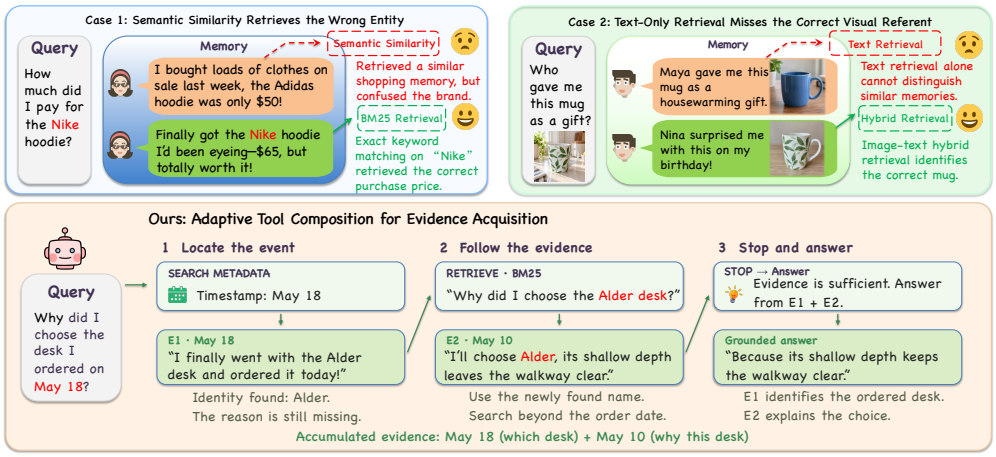


Figure 1: Illustrative cases motivating adaptive multimodal memory access. Top: failures of semantic-only and text-only retrieval, contrasted with successful BM25 and hybrid retrieval, respectively. Bottom: evidence-guided composition of metadata search, targeted retrieval, and stopping.

relevant yet still insufficient to support a correct answer to the query. As illustrated in Figure 1 (top), different queries may require different retrieval mechanisms, such as lexical matching or multimodal retrieval, highlighting the need for adaptive access to heterogeneous tools.

A complementary line of work explores iterative agentic retrieval, in which intermediate evidence is used to adapt subsequent memory access decisions. For example, Yan et al. (2025) and Du et al. (2025) interleave retrieval with reflection and evidence assessment, while Tian et al. (2026) extend tool-based planning to multimodal retrieval and analysis. More recent approaches such as Yu et al. (2026b) optimize memory operations with GRPO. Together, these methods make memory access more adaptive, allowing agents to revise their retrieval strategy as evidence accumulates. The remaining challenge is how to supervise these adaptive decisions according to the evidence currently available. A terminal answer reward indicates eventual success but does not explicitly identify which evidence is missing at an intermediate state. For memory access, effective guidance should address two questions at each step: whether the current evidence is sufficient and what remains missing. This motivates privileged feedback that is reconstructed from the evolving evidence state.

Therefore, we propose a novel framework for constructing and learning agentic access to long-term multimodal memory. We formulate memory access as a sequential decision process, with a unified tool-calling interface supporting heterogeneous retrieval, metadata search, event context expansion and source image inspection. The framework does not require elaborate memory construction or specialized memory structures, and operates directly over an external memory store through the unified tool interface. Instead of accumulating the complete history of tool observations in the context, our framework maintains a compact short-term memory (STM), an explicit evidence workspace that is updated after each tool call through reranking, merging, and deduplication. This keeps the model context compact while mitigating interference from irrelevant memory content. The resulting STM, together with the query and action history, defines the current decision state, from which the memory access policy selects the next action. While this framework provides the action space and evidence state required for adaptive memory access, the central challenge is how to supervise retrieval decisions according to the evidence currently available. Therefore, we introduce state-level on-policy self-distillation (OPSD). At each intermediate state of student rollouts, a verifier compares the STM with the reference answer and dynamically constructs privileged information (PI) characterizing evidence sufficiency and remaining information gaps. A frozen counterpart of the policy model serves as the teacher and is conditioned on the same state augmented with the PI. By exposing both evidence sufficiency and unresolved information gaps, the PI guides the teacher toward either terminating retrieval or acquiring the missing evidence. The student's action preferences are then aligned with those of the privileged teacher through a reverse-KL objective. This provides dense, state-specific guidance for tool selection based on the current evidence and remaining information gaps. As illustrated in Figure 1 (bottom), the agent first resolves the desk identity through metadata search and then uses the updated evidence to guide a BM25 search for the missing rationale. This demonstrates how the next retrieval action adapts to the evolving STM.

Despite recent progress in multimodal memory benchmarks, training data that capture long-horizon multimodal interactions remain limited. Therefore, we construct MemVista, a long-term multimodal memory dataset featuring multi-session conversations and temporally evolving facts. GPT-Image-2 is employed to generate and edit high-quality images, promoting close alignment between visual content and the events described in the conversations. Our contributions are summarized as follows:

- We introduce a state-level OPSD formulation for sequential evidence acquisition, where privileged feedback is dynamically constructed from the evolving STM as diagnoses of evidence sufficiency and unresolved information gaps. This design provides state-specific supervision for both retrieval termination and subsequent evidence acquisition.
- A unified agentic framework is developed for long-term multimodal memory access, providing heterogeneous memory-access operations together with a compact STM that consolidates intermediate evidence for sequential decision making.
- We construct *MemVista*, a training dataset of long-horizon conversations with temporally evolving facts, strong text–image alignment, and visual consistency.

2 Related Work

Long-Term Agent Memory. Research on long-term agent memory emphasizes both memory organization and adaptive access. MemGPT (Packer et al., 2023) and MemoryOS (Kang et al., 2025) adopt hierarchical memory management, while MIRIX (Wang & Chen, 2025) organizes multimodal experiences into specialized components. SimpleMem (Liu et al., 2026b) and V-Mem (Kang et al., 2026) improve retrieval through semantic compression, multi-view indexing, and modality-aware routing. Complementary agentic approaches, including GAM (Yan et al., 2025) and MemR^3 (Du et al., 2025), interleave retrieval with reflection and evidence assessment, while AdaMM (Tian et al., 2026) extends tool-based planning to multimodal memory. AgeMem (Yu et al., 2026b) further learns memory operations through reinforcement learning. Alongside these methods, LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2024) evaluate long-term conversational memory, while Mem-Gallery (Bei et al., 2026) and MemEye (Guo et al., 2026) extend evaluation to multimodal memory, including temporal reasoning and fine-grained visual evidence. However, most existing approaches rely primarily on textual memory representations or predetermined retrieval procedures, limiting their ability to accommodate the complexity of multimodal memory access. Our framework provides unified access to heterogeneous memory sources and a compact STM for evidence consolidation, while state-level OPSD leverages dynamically reconstructed privileged information to supervise retrieval decisions according to evidence sufficiency and unresolved information gaps.

On-Policy Distillation and Self-Distillation. On-policy distillation (OPD) provides teacher feedback on student-generated sequences, addressing the distribution mismatch associated with offline demonstrations, as studied in GKD (Agarwal et al., 2024). On-policy self-distillation obtains supervision from the same underlying model under more informative conditioning: Self-Distilled Reasoner (Zhao et al., 2026) uses privileged solution information, while SDPO (Hübotter et al., 2026) converts textual feedback into dense token-level supervision. For multi-step agents, TurnOPD (Zhou et al., 2026b) balances supervision across turns and adapts rollout depth, whereas MemOPD (Liu et al., 2026c) preserves teacher–student context alignment under memory compression. Unlike prior OPD and OPSD approaches, we dynamically construct privileged diagnoses at intermediate states of the student rollout. By characterizing evidence sufficiency and unresolved information gaps, they provide state-dependent context that guides the teacher toward appropriate subsequent decisions.

3 Method

3.1 Problem Formulation

Long-term multimodal memory access is formulated as a sequential retrieval problem. Let $\mathcal{M}$ denote an external long-term memory consisting of multi-session conversational events, with each event potentially comprising text, images, timestamps, and associated metadata. Given a user query q and an optional query image I_q, the objective is to retrieve and consolidate sufficient task-relevant evidence from $\mathcal{M}$ through a sequence of memory access operations.

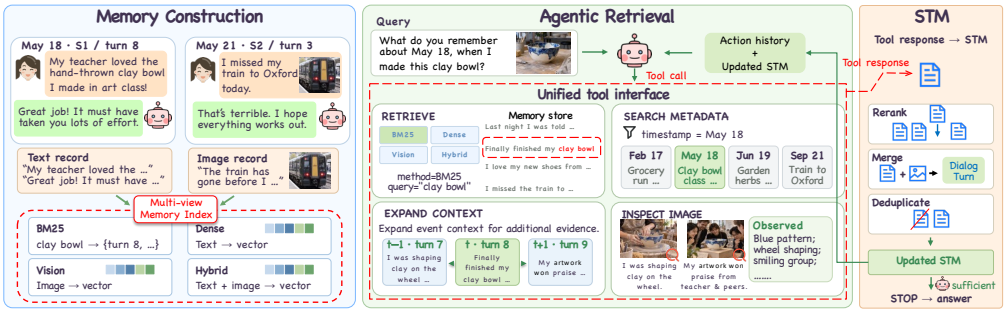


Figure 2: Overview of the agentic multimodal memory access framework. Multi-view memory indices support heterogeneous retrieval tools, while retrieved evidence is iteratively reranked, merged, and deduplicated into a compact short-term memory (STM) for subsequent retrieval decisions.

Rather than conditioning the model on the entire long-term memory or the full interaction history, a compact short-term memory STM_t is maintained to consolidate the evidence acquired up to step t. The corresponding decision state is defined as

$$s_t = (q, I_q, H_t, \mathrm{STM}_t)\,, \tag{1}$$

where $H_t = (a_1, \ldots, a_{t-1})$ denotes the history of previously executed actions. s_t is used to denote the observable state available to the policy. The policy selects the next action as

$$a_t \sim \pi_\theta(\cdot \mid s_t). \tag{2}$$

The action space comprises heterogeneous memory access operations, including textual and multimodal retrieval, metadata search, event-context expansion, and source-image inspection. Each action is executed against the external memory and returns an observation o_{t+1}. The STM is then updated by an evidence consolidation operator $\mathcal{U}$:

$$\mathrm{STM}_{t+1} = \mathcal{U}(\mathrm{STM}_t, o_{t+1})\,, \tag{3}$$

where $\mathcal{U}$ reranks, merges, and deduplicates newly acquired evidence to retain a compact set of relevant information. This enables each memory-access decision to be conditioned jointly on the original query and the evidence accumulated through preceding interactions.

The interaction terminates when the policy emits STOP or reaches the maximum action budget. A separate answer model then generates the final response conditioned on the query and the final STM:

$$\hat{y} = f_{\mathrm{ans}}\,(q, I_q, \mathrm{STM}_T)\,. \tag{4}$$

Since the final answer depends on the evidence retained in the STM, effective memory access requires a policy that determines when retrieval should terminate and what evidence should be acquired next. The central learning objective is therefore to make evidence-dependent memory-access decisions based on the evolving STM.

3.2 Agentic Multimodal Memory Access Framework

The proposed framework introduces an adaptive paradigm for long-term multimodal memory access, in which heterogeneous retrieval operations are dynamically composed according to the current evidence state rather than executed through a predefined pipeline. A unified tool interface enables complementary memory access, while a compact STM consolidates intermediate retrieval results to support subsequent decisions.

As illustrated in Figure 2, long-term multimodal memory access is built upon a multi-view external memory store. Each conversational turn is preserved as a multimodal event containing textual content, associated images, timestamps, and other metadata, while multiple indices are constructed to support complementary access patterns. Specifically, lexical (`bm25`) and semantic (`dense`) indices capture exact lexical and semantic associations, whereas visual (`vision`) and multimodal (`hybrid`) indices enable retrieval based on image and multimodal cues.

On top of this memory store, a unified agentic interface supports adaptive composition of heterogeneous memory-access operations. Rather than following a predetermined retrieval pipeline, each

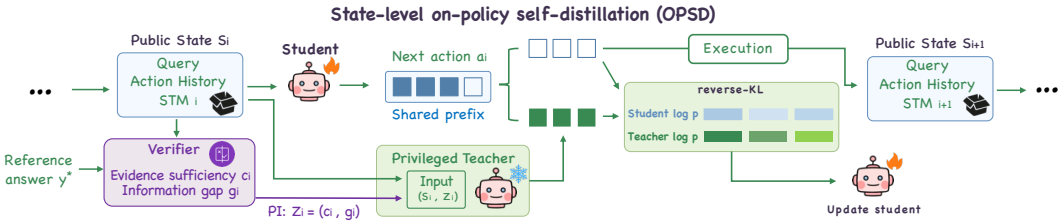


Figure 3: Overview of state-level on-policy self-distillation (OPSD). At each student state, a verifier constructs privileged information (PI) from the current STM. A frozen teacher conditioned on the PI provides action preferences, which are then distilled into the student via reverse-KL.

operation is selected according to the query and the evidence accumulated during execution. RETRIEVE exposes the lexical, semantic, visual, and multimodal retrieval modes described above. SEARCHMETADATA locates memories through structured attributes such as timestamps and modalities, while EXPANDCONTEXT recovers neighboring turns around an identified event to restore local conversational context and evolving facts. To limit visual redundancy and noise in the STM, retrieved images are represented by textual captions by default. When finer visual evidence is required, INSPECTIMAGE examines the source image with respect to the original query and returns a query-focused textual description of the relevant visual details. Together, these tools provide complementary access paths that can be adaptively composed across multiple retrieval steps.

Evidence returned by the tools is subsequently consolidated into a compact short-term memory (STM). Retrieved items are first mapped back to their corresponding conversational turns, after which are reranked, merged, and deduplicated before being incorporated into the STM. This turn-level consolidation preserves coherent textual, visual, and temporal context while suppressing redundant or irrelevant memory content. The updated STM then serves as the evidence state for the next retrieval step. This interaction proceeds iteratively until the policy emits STOP, after which the final STM is passed to a separate fixed answer model for response generation. Further implementation details of the memory access framework are provided in Appendix B.

3.3 STATE-LEVEL ON-POLICY SELF-DISTILLATION

The proposed framework provides the action space and evidence state required for adaptive memory access, but effective tool selection still requires learning how to act on the evolving STM. Learning such a memory access policy is challenging because the optimal action depends on the evidence available at each intermediate state. A final answer reward provides only coarse feedback for an entire retrieval trajectory, while offline teacher demonstrations are limited to a fixed state distribution and may not provide effective supervision for the states visited by the evolving student policy. These limitations call for supervision that is aligned with student-visited states and explicitly grounded in the evidence available at each state. We therefore introduce state-level OPSD, in which privileged supervision is organized around two evidence-dependent factors: evidence sufficiency and unresolved information gaps. As illustrated in Figure 3, these diagnoses are provided to a frozen teacher to induce state-specific action preferences, which are subsequently distilled into the student. Reconstructing the diagnoses at each student-visited state allows the supervision to reflect the evolving evidence state throughout the rollout. This state-wise construction also makes OPSD compatible with the non-prefix-preserving state updates induced by STM consolidation.

Dynamic privileged information. The privileged information is designed to expose the two decision factors required for state-aware memory access: whether retrieval should terminate and, if not, what evidence remains to be acquired. For each student-visited state $s_t = (q, I_q, H_t, \mathrm{STM}_t)$, a verifier compares the current STM against the reference answer y^* and constructs

$$z_t = (c_t, g_t) = \mathcal{V}\left(q, I_q, \mathrm{STM}_t, y^*\right), \tag{5}$$

where c_t indicates evidence sufficiency and g_t characterizes the remaining information gap. The reference answer is used only by the verifier to construct z_t and is not exposed to either the student or the teacher. This separation reduces the risk of direct PI leakage (Yang et al., 2026) by preventing the teacher from being directly conditioned on the reference answer during distillation.

Conditioning on z_t provides the teacher with explicit guidance on evidence sufficiency and the unresolved information gap, thereby inducing a diagnosis-informed action distribution for the current state. Since the student observes only the public state s_t, distillation encourages its action preferences to align with this evidence-dependent decision rule based on the evolving STM. z_t is reconstructed as the STM changes, keeping teacher supervision aligned with the evidence state.

Diagnosis-conditioned privileged teacher. A frozen counterpart of the policy model serves as the privileged teacher. At each student-visited state, both the student and teacher share the same public state s_t, while the teacher is additionally conditioned on z_t. The resulting teacher distribution therefore provides diagnosis-informed preferences over subsequent memory-access actions. These preferences are distilled into the student through a sampled-token reverse-KL objective:

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{\substack{s_t \sim d^{\pi_\theta}, \\ x_t \sim \pi_\theta(\cdot|s_t)}} \left[\frac{1}{L_t} \sum_{j=1}^{L_t} \left(\log \pi_\theta(x_{t,j} \mid s_t, x_{t,<j}) - \log \pi_{\bar{\theta}}(x_{t,j} \mid s_t, z_t, x_{t,<j}) \right) \right], \tag{6}$$

where d^{π_θ} denotes the state distribution induced by the current student policy, $x_t = (x_{t,1}, \ldots, x_{t,L_t})$ is a tool-call sequence of length L_t sampled from $\pi_\theta(\cdot \mid s_t)$, and π_θ and $\pi_{\bar{\theta}}$ denote the student and frozen privileged teacher policies, respectively. For each sampled sequence, the student and teacher evaluate the same action prefixes $x_{t,<j}$, with the teacher additionally conditioned on z_t. Thus, diagnosis-informed teacher preferences are evaluated on the states and action prefixes induced by the current student policy. Equation (6) therefore defines a length-normalized sampled-token surrogate for reverse-KL alignment between the student and privileged teacher distributions. Further implementation details of state-level OPSD are provided in Appendix C.

3.4 MemVista Construction Pipeline

The proposed framework and state-level OPSD require training data that can support sequential, evidence-dependent memory access. While recent work has introduced several benchmarks for evaluating multimodal memory, training data tailored to sequential memory access remain limited. Addressing this gap requires long-horizon multimodal conversations paired with temporally grounded queries and traceable supporting evidence for supervising intermediate retrieval states. To this end, we construct *MemVista*, a training dataset specifically designed for agentic long-term multimodal memory access. MemVista contains multi-session multimodal conversations with temporally evolving facts, together with diverse memory-oriented questions and explicit supporting evidence. Its construction pipeline further enforces cross-session consistency and evidence traceability, providing the supervision required for state-level policy learning.

Long-horizon event planning and multimodal generation. Each memory instance is first organized around a persistent character profile and a temporally ordered sequence of events, specifying factual updates, conversational context, and the visual content associated with each event. Existing multimodal memory benchmarks such as Mem-Gallery and MemEye typically pair generated conversations with predominantly real images, which can result in weak text-image alignment and inconsistent visual entities across sessions. MemVista instead generates visual observations jointly with the conversational events. Each event is accompanied by a set of dedicated seed prompts, which condition GPT-Image-2 to synthesize a corresponding visual observation. When entities or scenes evolve over time, image editing is applied to preserve visual identity while reflecting the intended changes. This design yields tightly aligned text-image conversations with substantially stronger cross-session visual consistency. To ensure that supervision is grounded in the generated memory rather than the underlying plan, only information explicitly realized in the dialogue or verified from the corresponding visual observation is retained as an *observed fact*. These observations are recorded in a temporal ledger that tracks fact validity and updates across sessions, enabling temporally grounded QA construction with traceable supporting evidence.

QA construction and evidence annotation. Based on the observed facts, memory questions are constructed to cover cross-session recall, temporal reasoning, multimodal grounding, and fact evolution. Each question is paired with a reference answer and traceable supporting evidence from the corresponding conversational events and images. The resulting instances are exported as interactive memory environments containing the public text–image history and queries, while reference

answers and evidence annotations are retained as private training-time information. This separation supports both multi-step memory-access rollouts and the dynamic construction of privileged information required by state-level OPSD.

Two-stage data filtering. To ensure reliable supervision, all QA instances undergo two stages of quality filtering. First, when the annotated supporting memory is provided, the reference answer must be recovered correctly, eliminating ambiguous or intrinsically under-specified questions. Second, the same question is evaluated with the annotated supporting memory removed while all remaining conversational context is retained. The answer must then be incorrect or appropriately rejected. This second criterion removes instances affected by evidence leakage or questions answerable from the model's internal knowledge alone. Only instances satisfying both criteria are retained. Further details of MemVista construction are provided in Appendix D.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. Experiments are conducted on two complementary long-term multimodal memory benchmarks. Mem-Gallery (Bei et al., 2026) contains 1,711 questions over long-horizon multimodal conversations and evaluates test-time learning (TTL), visual-centric reasoning (VR), visual-centric search (VS), answer refusal (AR), conflict detection (CD), knowledge resolution (KR), multi-entity reasoning (MR), factual retrieval (FR), and temporal reasoning (TR). MemEye (Guo et al., 2026) places greater emphasis on visual memory, ranging from scene-level information to fine-grained visual details and evolving visual states. Together, they evaluate both general multimodal memory reasoning and fine-grained visual evidence acquisition.

Baselines. Comparisons are conducted against Omni-SimpleMem (Liu et al., 2026a), V-Mem (Kang et al., 2026), and MemR^3 (Du et al., 2025), covering representative retrieval-based and agentic memory systems. Two trained variants of the proposed framework are additionally evaluated: *Ours (GRPO)*, optimized with trajectory-level rewards, and *Ours (OPSD)*, trained with the proposed state-level privileged supervision.

Implementation Details. Qwen3.5-4B (Qwen Team, 2026) is used as the policy backbone for both GRPO and OPSD, with the same initialization, memory-access tools, and action budget. Qwen3.5-9B provides rewards for GRPO and constructs privileged information for OPSD, while a frozen Qwen3.5-4B counterpart serves as the OPSD teacher. External baselines follow their released implementations and recommended settings. To control for answer generation capability, all methods are evaluated under two shared answer and judge settings, using either Qwen3.5-9B or GPT-5.5 for both components. Further implementation and training details are provided in Appendix F.

4.2 MAIN RESULTS

Results on Mem-Gallery. Table 1 shows that the proposed method consistently achieves the best overall LLM-judge performance under both answer-and-judge settings, reaching 0.852 with Qwen3.5-9B and 0.893 with GPT-5.5. It surpasses V-Mem, the strongest external baseline, and also outperforms MemR^3, a representative agentic baseline, highlighting the importance of effective policy supervision for agentic memory access. Compared with the GRPO variant, OPSD achieves higher overall performance on Mem-Gallery. This improvement is consistent with state-level supervision adapting to the evolving STM, allowing the policy to target unresolved evidence gaps and terminate once sufficient evidence has been acquired.

Results on MemEye. Table 3 presents the results on MemEye, where both learned variants of the proposed framework substantially outperform the external memory baselines. Under the Qwen3.5-9B setting, GRPO achieves the best overall performance, while OPSD remains competitive with a Judge score of 0.582 and an accuracy of 49.33%. Under the GPT-5.5 setting, OPSD performs best on both metrics, reaching a Judge score of 0.650 and an accuracy of 55.80%, with a clear margin over V-Mem, the strongest external baseline. These results highlight the effectiveness of adaptive multimodal tool composition for the fine-grained visual evidence acquisition required by MemEye.

Table 1: LLM-judge scores on Mem-Gallery under Qwen3.5-9B and GPT-5.5 answer-and-judge settings. External baselines are evaluated using their original memory access mechanisms, while *Ours (GRPO)* and *Ours (OPSD)* share the same policy backbone, memory-access tools, and action budget. Higher scores indicate better performance.

System	TTL	VR	VS	AR	CD	KR	MR	FR	TR	Overall
Qwen3.5-9B Answer/Judge										
Omni-SimpleMem	0.677	0.463	0.675	0.870	0.531	0.648	0.813	0.806	0.732	0.704
MemR^3	0.700	0.494	0.176	0.807	0.728	0.617	0.757	0.845	0.715	0.621
V-Mem	0.766	0.569	0.848	0.973	0.543	0.673	0.845	0.845	0.658	0.780
Ours (GRPO)	0.862	0.780	0.877	0.596	0.618	0.815	0.909	0.952	0.822	<u>0.828</u>
Ours (OPSD)	0.944	0.736	0.917	0.780	0.605	0.759	0.900	0.879	0.805	**0.852**
GPT-5.5 Answer/Judge										
Omni-SimpleMem	0.684	0.658	0.845	0.826	0.704	0.741	0.898	0.879	0.878	0.794
MemR^3	0.761	0.514	0.177	0.796	0.716	0.722	0.837	0.861	0.776	0.654
V-Mem	0.907	0.827	0.926	0.828	0.857	0.801	0.905	0.898	0.854	<u>0.881</u>
Ours (GRPO)	0.921	0.879	0.890	0.685	0.889	0.883	0.915	0.938	0.828	0.877
Ours (OPSD)	0.935	0.897	0.913	0.788	0.877	0.821	0.927	0.920	0.833	**0.893**

Table 2: Qualitative case study on Mem-Gallery. The table reports the original tool-call trajectories produced by the model for two representative examples, together with brief analysis of the behavior.

Query	Tool-Call Trajectory	Analysis
Which image corresponds to the Botanical Floral Shift Dress that Maria selected?	`RETRIEVE` `{"method":"hybrid", "query":"Botanical Floral Shift Dress Maria selected"}` → `INSPECT_IMAGE` → `STOP`	**Correct**. Hybrid retrieval identifies the relevant candidate memories, while INSPECTIMAGE provides targeted visual verification before termination.
Please search for all photos of the same plant species as this image shared in the conversation.	`RETRIEVE` `{"method":"hybrid", "query":"purple flowers plant species"}` → `RETRIEVE` `{"method":"bm25", "query":"lisianthus purple flower plant species"}` → `STOP`	**Correct**. The query does not explicitly specify the plant species. Hybrid retrieval first identifies the plant as lisianthus, after which the inferred species name is incorporated into a BM25 query to retrieve matching photos.

The competitive performance of OPSD across both evaluation settings further suggests that state-level supervision provides an effective learning signal for evidence acquisition beyond conventional trajectory-level reward optimization. Additional analysis of evidence coverage in the final evidence set is provided in Appendix G.1. Tool usage statistics are further reported in Appendix G.2 to examine how the learned policies compose different operations across the two benchmarks.

4.3 CASE STUDY

To further examine the retrieval behavior learned by the proposed framework, Table 2 presents two representative examples from Mem-Gallery. In the first case, hybrid retrieval identifies candidate memories associated with the selected dress, after which INSPECTIMAGE is invoked to verify the relevant visual evidence before termination. This demonstrates that visual inspection can be introduced when the retrieved evidence is insufficient for resolving a visually grounded query. In the second case, the plant species is not explicitly specified in the query. Hybrid retrieval first resolves the species as *lisianthus*, and this newly acquired information is subsequently incorporated into a BM25 query to retrieve the corresponding photos. The retrieval sequence therefore evolves with the accumulated evidence rather than being determined solely by the original query.

Together, these examples illustrate how the framework composes complementary memory access operations according to the current evidence state. In particular, intermediate retrieval results can either trigger a more targeted modality-specific operation or provide new semantic cues for subse-

Table 3: Results on MemEye. Judge denotes the LLM-based evaluation score of the generated answers, and Acc. (%) denotes the proportion of strictly correct samples. Higher is better.

	Qwen3.5-9B		*GPT-5.5*	
System	Judge	Acc. (%)	Judge	Acc. (%)
Omni-SimpleMem	0.303	24.80	0.403	32.88
MemR^3	0.321	26.42	0.313	25.07
V-Mem	0.340	26.95	0.514	43.94
Ours (GRPO)	**0.592**	**50.67**	0.607	53.64
Ours (OPSD)	0.582	49.33	**0.650**	**55.80**

Table 4: Ablation results on Mem-Gallery and MemEye. Avg Actions denotes the average tool-calling rounds per query.

Method	Acc. (%)	Judge	Avg Actions
Mem-Gallery			
Base Agent	68.67	0.799	6.45
+ Reference-Trajectory PI	71.71	0.835	3.44
+ Reference-Trajectory PI + STM	72.53	0.848	4.31
Ours (State-level PI + STM)	**73.93**	**0.852**	**2.89**
MemEye			
Base Agent	44.23	0.528	7.81
+ Reference-Trajectory PI	44.16	0.532	6.08
+ Reference-Trajectory PI + STM	46.01	0.548	6.95
Ours (State-level PI + STM)	**49.33**	**0.582**	**5.66**

quent search, enabling the retrieval policy to progressively refine the evidence acquisition process. Additional qualitative examples, including representative failure cases, are provided in Appendix E.

4.4 Ablations

Table 4 presents four controlled variants designed to isolate the contributions of the STM mechanism and privileged-information construction. *Base Agent* denotes the base Qwen3.5-4B policy without STM or policy optimization. + *Reference-Trajectory PI* applies OPSD by providing the frozen teacher with a correct execution trajectory as privileged information, while conditioning the teacher on the current state encountered by the student. The reference trajectory remains fixed throughout the student rollout. + *Reference-Trajectory PI + STM* further introduces STM-based evidence consolidation through reranking, merging, and deduplication. Finally, *Ours (State-level PI + STM)* replaces the fixed reference trajectory with dynamic diagnoses of evidence sufficiency and unresolved information gaps constructed at each student-visited state. Further details on the construction and use of Reference-Trajectory PI are provided in Appendix C.

Effect of STM. Under Reference-Trajectory PI, introducing STM improves accuracy from 71.71% to 72.53% on Mem-Gallery and from 44.16% to 46.01% on MemEye, with corresponding improvements in LLM-judge scores. This supports the use of a consolidated evidence state for subsequent memory-access decisions.

Effect of state-level PI. With STM retained, replacing Reference-Trajectory PI with state-level PI further improves accuracy by 1.40 and 3.32 percentage points on Mem-Gallery and MemEye, respectively. More importantly, average action counts decrease by 32.9% and 18.6%, yielding the best accuracy and fewest actions among the evaluated ablation variants on both benchmarks. By reconstructing PI from the current STM at each student-visited state, teacher supervision adapts to the student's actual retrieval progress, directing the policy toward specific missing evidence when necessary and toward termination once the evidence becomes sufficient. This state-aligned guidance improves evidence acquisition while avoiding unnecessary retrieval steps.

5 Conclusion

We presented an agentic framework for adaptive access to long-term multimodal memory, combining heterogeneous memory-access operations with an incrementally consolidated STM. Our central contribution is state-level OPSD, which learns evidence-dependent memory-access policies using dynamic privileged supervision based on evidence sufficiency and unresolved information gaps. These diagnoses induce state-specific action preferences in the privileged teacher, which are distilled into a student that relies only on the public state at inference time. We further constructed MemVista, a training dataset featuring long-horizon multimodal conversations, temporally evolving facts, and traceable supporting evidence. Experiments on Mem-Gallery and MemEye demonstrate strong performance over existing memory systems, while controlled ablations show that replacing fixed reference-trajectory PI with state-dependent diagnoses improves accuracy while reducing memory-access actions. Together, these results demonstrate the value of aligning privileged supervision with the evolving evidence state for adaptive multimodal memory access.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript, as well as for data generation in the construction of MemVista. Specifically, GPT-5.5 was used for language-based stages of data construction, including event planning, dialogue generation, and QA construction, while GPT-Image-2 was used for image generation and editing. We did not use generative AI tools to develop the proposed method or framework, formulate the experimental design, interpret experimental results, or draw the conclusions of this work. All AI-assisted content, including generated data and manuscript edits, was reviewed and verified by the authors, and generated data were further processed and filtered following the quality-control procedures described in the paper and supplementary material. We take full responsibility for all AI-assisted content, including manuscript edits and generated data, as well as for the final scientific content of this work.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of private user data. MemVista consists of synthetic multi-session conversations and images generated or edited with GPT-Image-2. This avoids direct exposure of personal communication records. Dataset releases will include documentation of the construction and filtering procedures.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. The agentic multimodal memory access framework and state-level OPSD are formally described in Sections 3.2 and 3.3, including the memory access interface, STM construction, privileged information, and distillation objective. Experimental settings, backbone models, evaluation protocols, and baseline configurations are documented in Section 4.1. The supplementary material provides further implementation and training details, including tool implementations, hyperparameters, prompts, and additional case studies. The complete MemVista construction pipeline, including multimodal generation, QA construction, evidence annotation, and two-stage data filtering, is also documented in the supplementary material. Code for reproducing the main experiments will be included in the supplementary materials.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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.

Xingbo Du, Loka Li, Duzhen Zhang, and Le Song. MemR^3: Memory retrieval via reflective reasoning for llm agents. *arXiv preprint arXiv:2512.20237*, 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.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, et al. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *arXiv preprint arXiv:2606.15912*, 2026a.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026b.

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*, 2026a.

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

Zhiyuan Liu, Tinghong Ye, Chenghao Liu, Yizhuo Li, and Songfang Huang. Memopd: On-policy distillation through memory state alignment for long-horizon agents. *arXiv preprint arXiv:2608.07068*, 2026c.

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

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity. *arXiv preprint arXiv:2606.26091*, 2026.

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. On memory construction and retrieval for personalized conversational agents. *arXiv preprint arXiv:2502.05589*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

Alina Shutova, Alexandra Olenina, Ivan Vinogradov, and Anton Sinitsin. Evaluating memory structure in llm agents. *arXiv preprint arXiv:2602.11243*, 2026.

Qitai Tan, Zefang Zong, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2606.27814*, 2026.

Zhoujin Tian, Yao Tian, Hao Zhang, Cheng Chen, Yakun Li, Lei Zhang, and Xiaofang Zhou. Beyond retrieval: Analytic memory for multimodal agents. *arXiv preprint arXiv:2607.29440*, 2026.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, et al. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*, 2026.

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

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

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

BY Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026a.

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, 2026b.

Guibin Zhang, Jiayang Lyu, Ran Sun, Xinlei Yu, Haoyu Zhao, Qibing Ren, and Shuicheng Yan. Latent on-policy self-distillation. *arXiv preprint arXiv:2608.13040*, 2026a.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026b.

Yunxiang Zhang, Yiheng Li, Ali Payani, and Lu Wang. Adamem: Test-time adaptive memory for language agents. *arXiv preprint arXiv:2606.05684*, 2026c.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 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.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Sage-opd: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026a.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. Turnopd: Making on-policy distillation turn-aware for efficient long-horizon agent training. *arXiv preprint arXiv:2607.05804*, 2026b.

A Extended Related Work

Long-term and multimodal memory. A broad body of work studies long-term memory for language models and agents, spanning external databases, persistent conversational memory, hierarchical memory management, structured memory construction, and adaptive retrieval (Hu et al., 2023; Zhong et al., 2024; Packer et al., 2023; Maharana et al., 2024; Wu et al., 2024; Xu et al., 2026; Chhikara et al., 2025; Pan et al., 2025; Kang et al., 2025). Recent systems further explore richer memory organization and multimodal memory representations (Wang & Chen, 2025; Liu et al., 2026b;a; Kang et al., 2026), while agentic approaches increasingly allow memory operations to be selected or refined during interaction (Yan et al., 2025; Du et al., 2025; Zhang et al., 2026c; Tian et al., 2026; Yu et al., 2026b). In parallel, a growing set of benchmarks evaluates long-term recall, temporal reasoning, memory organization, and multimodal evidence use (Maharana et al., 2024; Wu et al., 2024; Bei et al., 2026; Guo et al., 2026; Shutova et al., 2026; Li et al., 2026b).

Agentic on-policy distillation. On-policy distillation provides teacher supervision on student-generated trajectories, reducing the distribution mismatch associated with distillation on fixed demonstrations (Agarwal et al., 2024). This paradigm has recently been extended through self-distillation and privileged teacher conditioning (Zhao et al., 2026; Wang et al., 2026), as well as step-aware supervision for long-horizon agents (Zhong et al., 2026; Zhang et al., 2026b; Li et al., 2026a; Zhou et al., 2026a; Tan et al., 2026; Zhou et al., 2026b). Other recent studies investigate supervision reliability, privileged information, adaptive distillation targets, and stability in OPSD (Yu et al., 2026a; Shen et al., 2026; Nicolicioiu et al., 2026; Zhang et al., 2026a). MemOPD (Liu et al., 2026c) further studies state alignment when OPD is combined with compressed agent memory, extending on-policy distillation to long-horizon memory-based interaction.

B Additional Method Details

Unified tool interface. The five memory access operations introduced in the main paper share a unified function calling interface with typed arguments. RETRIEVE supports four modes (`bm25`, `dense`, `vision`, and `hybrid`) with `top_k` $\in [1, 50]$ and an optional rewritten query; EXPANDCONTEXT uses a window $w \in \{1, 2, 3\}$. Other tools expose only their task-specific arguments. Rewritten queries apply only to the corresponding retrieval call, while the original query and optional query image remain available throughout the interaction. Representative raw tool-call outputs are provided in Table 5.

STM construction and evidence consolidation. The discovery operations RETRIEVE, SEARCHMETADATA, and EXPANDCONTEXT share the same evidence consolidation procedure. Retrieved records are first mapped to their corresponding conversational turns, so that textual content, images, and metadata from the same turn are represented as a single event. Newly discovered events are then merged with the existing candidate pool and deduplicated at the event level. For subsequent updates, candidate rankings are combined using reciprocal rank fusion,

$$R(e) = \sum_{r \in \mathcal{R}(e)} \frac{1}{60 + r}, \tag{7}$$

where $\mathcal{R}(e)$ denotes the ranks assigned to event e by the contributing candidate lists. The candidate pool is capped at 24 events in the reported configuration.

Retrieval implementations. All RETRIEVE modes first compute scores at the record level. Records belonging to the same conversational turn are aggregated by maximum score, after which scores are propagated to immediately adjacent turns within the same session:

$$s_t = \max_{r \in t} s(r), \qquad \widetilde{s}_t = \max\{s_t, 0.9 s_{t-1}, 0.9 s_{t+1}\}. \tag{8}$$

This propagation modifies only the retrieval ranking and is distinct from the explicit EXPANDCONTEXT operation. The top-k conversational turns are passed to candidate consolidation. Therefore, k specifies the retrieval depth before STM construction rather than the final number of evidence items.

BM25 performs lexical retrieval over the concatenation of record summaries and textual content, including stored descriptions associated with image records. We use $k_1 = 1.5$ and $b = 0.75$. Text

Table 5: Raw tool-call outputs extracted directly from evaluation trajectories. Empty and null fields are omitted only for readability, while function names, tool-call IDs, and argument strings are preserved exactly as recorded in the original assistant messages.

Retrieval. Query: Which image shown in 2024-06-17 is an application of artificial intelligence in the field of education?


```
{"role":"assistant","tool_calls":[{
  "id":"chatcmpl-tool-99f869e5c678333d","type":"function",
  "function":{"name":"RETRIEVE",
    "arguments":"{\"method\": \"hybrid\", \"top_k\": 50, \"query\":
    \"2024-06-17 artificial intelligence education image\"}"}}]}
```


Metadata Search. Query: How many photos are related to vehicles in the conversation of 2024-07-21?


```
{"role":"assistant","tool_calls":[{
  "id":"chatcmpl-tool-aa56547e674f51ba","type":"function",
  "function":{"name":"SEARCH_METADATA",
    "arguments":"{\"field\": \"timestamp\", \"op\": \"contains\", \"
    value\": \"2024-07-21\"}"}}]}
```


Context Expansion. Query: How many blue cards does Player 1 hold immediately after their visible hand size changes from 6 to 5 for the first time?


```
{"role":"assistant","tool_calls":[{
  "id":"chatcmpl-tool-982f4e76f74d3ffb","type":"function",
  "function":{"name":"EXPAND_CONTEXT",
    "arguments":"{\"window\": 2}"}}]}
```


Image Inspection. Query: Which of the images uploaded in 2024-07-07 focus on a robotic arm?


```
{"role":"assistant","tool_calls":[{
  "id":"chatcmpl-tool-90af423b547f38f5","type":"function",
  "function":{"name":"INSPECT_EVIDENCE_IMAGE",
    "arguments":"{\"target\": \"current_evidence_images\", \"instruction
    \": \"answer_query_related_visual_details\"}"}}]}
```


Stop. Query: Which image shown in 2024-06-17 is an application of artificial intelligence in the field of education?


```
{"role":"assistant","tool_calls":[{
  "id":"chatcmpl-tool-aefb97ac17413969","type":"function",
  "function":{"name":"STOP","arguments":"{}"}}]}
```


is normalized to lowercase and tokenized into alphanumeric tokens and Chinese characters, with adjacent Chinese character bigrams additionally included. No stemming or stop-word removal is applied. BM25 does not use the query image and is primarily intended for exact names, entities, and lexical cues.

Dense retrieval uses `all-MiniLM-L6-v2` to construct 384-dimensional normalized text representations. Text records and image descriptions are encoded with masked mean pooling over the final hidden states. In addition to the complete query, up to six clause-level query views are constructed, and each record is scored by its maximum cosine similarity across these views. Dense retrieval therefore captures semantic correspondence and paraphrases but does not directly use image pixels.

Vision retrieval uses `SigLIP-Base-Patch16-384` with 768-dimensional normalized representations. When a query image is available, image-to-image similarity is computed between the query image and memory images. In this branch, the retrieval text does not affect visual ranking. Without a query image, the SigLIP text encoder is used for text-to-image retrieval. Retrieved image records are subsequently mapped back to their complete conversational turns.

Hybrid retrieval uses `gme-Qwen2-VL-2B-Instruct` to construct 1536-dimensional joint representations. Text records are encoded from their textual content, while image records jointly encode

the image pixels and associated text. A multimodal query jointly represents the retrieval text and query image when available. Similarity is computed by normalized inner product. Unlike VISION, query reformulation can therefore affect the ranking even when a query image is present. The primary hybrid implementation is based on this joint representation.

Metadata search and context expansion. SEARCHMETADATA performs an independent structured search over the external memory. Modality constraints support equality or inequality over `text` and `image`, while timestamp constraints support `eq`, `contains`, `before`, and `after`. Timestamp matching operates on stored conversational timestamps rather than temporal expressions mentioned inside the dialogue. Each metadata call specifies a single constraint and does not impose a persistent filter on subsequent retrieval operations.

Expand retrieved event context. EXPANDCONTEXT expands around all events currently contained in the STM. Given $w \in \{1, 2, 3\}$, up to w preceding and following turns are collected within the same conversational session and merged with the candidate pool. Expansion never crosses session boundaries. This operation is intended to recover local conversational information such as references, corrections, enumerations, and event transitions that may not be contained in an individually retrieved turn.

Source image inspection. Directly exposing all retrieved source images to the model would substantially increase the multimodal context and introduce irrelevant visual information. Images are therefore represented in the STM by textual descriptions by default. When finer visual evidence is required, INSPECTIMAGE examines previously uninspected source images associated with the current STM. For each eligible image, a vision-language model is conditioned on the original query, the same-turn conversational content, the source image, and, when available, the query image as a comparison reference. The model is instructed to extract only directly visible details relevant to the query rather than generate the final answer.

C ADDITIONAL DETAILS OF STATE-LEVEL OPSD

State-wise supervision construction. State-level OPSD constructs supervision independently at each decision state encountered by the student. Let $s_t = (q, I_q, H_t, \mathrm{STM}_t)$ denote the public state before the next memory access action is executed. The student samples an action from $\pi_\theta(\cdot \mid s_t)$ using only this public information. A separate verifier then evaluates the current evidence state against the private reference answer y^* and constructs the privileged information z_t. Importantly, the verifier diagnoses the evidence available at the current state rather than evaluating the completed retrieval trajectory. Consequently, z_t is recomputed after every STM update and may change as new evidence is acquired.

Verifier and privileged information. In the implementation, the conceptual privileged information $z_t = (c_t, g_t)$ introduced in the main paper is instantiated with a structured information gap

$$g_t = (\tau_t, r_t), \qquad z_t = (c_t, \tau_t, r_t), \tag{9}$$

where $c_t \in \{0, 1\}$ denotes evidence sufficiency, τ_t specifies the type of unresolved information, and r_t provides a short description of the remaining gap grounded in the current public state. The verifier receives the query, current STM and search progress, together with the reference answer as a private rubric. The reference answer is used only to construct z_t and is never exposed to either the student or the privileged teacher.

The verifier is instructed to determine whether every component required by the query is supported by the current evidence. For list and counting questions, sufficiency requires adequate coverage rather than a single matching example. For visual questions, source image inspection is required only when the visual attribute needed by the query is not already resolved by the available description or previous visual observation. For questions involving absence or missing information, evidence is considered sufficient only after multiple complementary search attempts provide adequate coverage. These criteria make the privileged information reflect the unresolved requirement of the current query rather than a generic relevance judgment.

Table 6: Types of unresolved information used by the state verifier. The category τ_t forms part of the information gap g_t and summarizes what remains unresolved in the current STM.

Type	Description
`none`	All evidence required by the query is available.
`no_public_evidence`	No relevant public evidence has been acquired.
`irrelevant_evidence`	Current evidence concerns an incorrect entity, event, topic, or modality.
`missing_metadata_constraint`	A required temporal or modality constraint has not been isolated.
`candidate_set_too_broad`	Retrieved candidates remain ambiguous or insufficiently focused.
`missing_neighbor_context`	A relevant event has been found but surrounding conversational context is missing.
`missing_temporal_order`	The required ordering, extremum, or temporal relation remains unresolved.
`missing_raw_visual_detail`	Relevant image evidence is available but necessary visual detail remains unresolved.
`incomplete_coverage`	Evidence covers only part of a required list or counting set.
`missing_factual_support`	A required fact, relation, attribute, or event is unsupported.

Diagnostic normalization and leakage control. Verifier outputs are parsed and normalized before being used for distillation. Logical consistency between c_t and τ_t is enforced. For example, a sufficient state must be paired with `none`, whereas an insufficient state must specify a non-empty gap type. Diagnostics that conflict with the observable state are corrected to a compatible category when possible. The textual explanation is additionally screened for potential leakage from the private reference answer and replaced by a more generic description when necessary. Verifier responses that cannot be parsed or safely normalized are excluded from state-level supervision rather than used as noisy training targets. These mechanisms provide practical safeguards against reference-answer leakage, although they should be regarded as heuristic checks rather than formal guarantees.

Privileged teacher conditioning. A frozen Qwen3.5-4B counterpart of the student policy serves as the privileged teacher. For each valid student state, the teacher receives the original query, public action history, current STM and search status, together with the sanitized privileged information z_t. The teacher does not receive the reference answer or annotated supporting evidence. Its available tool schemas are identical to those exposed to the student at the corresponding state.

The privileged information modifies the teacher's decision context rather than directly prescribing a target action. When c_t indicates sufficient evidence, the teacher is encouraged to prefer STOP; otherwise, the diagnosed information gap guides the teacher toward an available operation that can address the unresolved requirement. Importantly, tool arguments must still be grounded in public information from the query, action history, and current STM. Thus, privileged information determines *what remains unresolved* without revealing the missing answer itself.

On-policy teacher scoring. State-level OPSD does not require the privileged teacher to generate an independent retrieval trajectory. Instead, the action sampled by the student is reused as the sequence to be evaluated by both policies. For a student-sampled tool call $x_t = (x_{t,1}, \ldots, x_{t,L_t})$, the student probabilities are obtained from the original rollout, whereas the frozen teacher evaluates the same action tokens under its privileged context using teacher forcing:

$$\pi_\theta(x_{t,j} \mid s_t, x_{t,<j}), \qquad \pi_{\bar{\theta}}(x_{t,j} \mid s_t, z_t, x_{t,<j}). \tag{10}$$

The sampled-token reverse-KL objective described in the main paper is applied only to the tool-call tokens. This keeps supervision aligned with the action sequence and state distribution actually induced by the current student policy. For optimization, the teacher-minus-rollout-student log-probability difference on student-sampled tool-call tokens is detached and used as a token-level advantage in a PPO-style clipped surrogate objective. Gradients flow only through the current-to-rollout policy probability ratio, with the advantage and rollout log-probabilities held fixed during backpropagation.

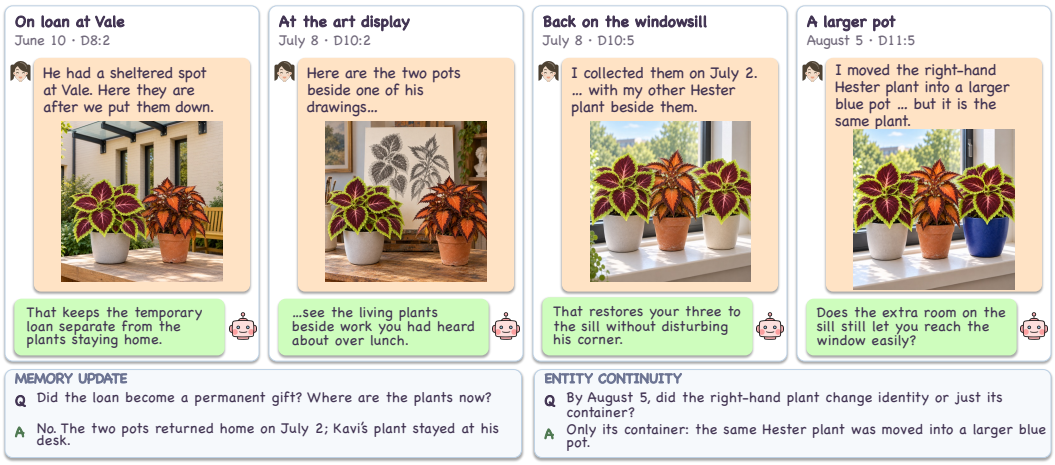


Figure 4: Representative MemVista example with temporally distributed multimodal events. The upper panels track the movement and modification of visually persistent entities across sessions, while the lower panels illustrate memory-update and entity-continuity questions grounded in the accumulated history.

Reference-Trajectory PI. For the Reference-Trajectory PI baseline, Qwen3.5-9B is first rolled out on each training query, and a correct memory-access trajectory is selected as the reference trajectory for that query. During OPSD training, this trajectory is kept fixed across all student-visited states and provided to the frozen teacher as static privileged information together with the current student state. The reference trajectory serves only as privileged context for the teacher rather than as a direct imitation target. In contrast, our state-level PI is reconstructed from the evolving STM at each student-visited state and therefore adapts to the student's current retrieval progress.

D MemVista Construction Details

D.1 Construction Pipeline

MemVista is constructed as a training corpus for agentic long-term multimodal memory access. Each instance contains a temporally extended multimodal history, memory-oriented questions, reference answers, and traceable supporting evidence. GPT-5.5 is used for the language-based stages of data construction, including event planning, dialogue generation, and QA construction, while GPT-Image-2 is used for image generation and editing. The overall pipeline consists of long-horizon event planning, event-aligned multimodal generation, observed-fact extraction, QA and evidence construction, and multi-stage quality control. Importantly, supervision is derived from the realized conversations and visual observations rather than directly from the private generation plan. Figure 4 provides a representative example of temporally evolving multimodal events with consistent visual entities across sessions.

Long-horizon event planning. Each memory instance is organized around a persistent character profile and a temporally ordered event sequence. Each event corresponds to a user-assistant exchange with associated temporal metadata and optional visual content. Factual updates are realized through later events, allowing identities, attributes, preferences, and relationships to evolve over time while preserving their earlier states.

Event-aligned multimodal generation. To improve text-image alignment and cross-session visual consistency, visual content is generated in coordination with the underlying event design rather than paired with independently sourced real images. Each visual event is assigned a semantic specification describing the entities, attributes, and relations that should be visually expressed, together with a set of dedicated seed prompts. GPT-Image-2 is then used to synthesize the corresponding image. When an entity or scene has appeared previously, the earlier image is used as a reference for

image editing, preserving visual identity while reflecting the intended event-level changes. Generated images are subsequently associated with the corresponding conversational events.

The dialogue is generated sequentially according to the event timeline. User and assistant turns are conditioned on the conversational history and event-local information, while designated visual attributes can be withheld from the text to prevent multimodal questions from becoming answerable solely from textual descriptions.

D.2 Observed Facts and Temporal Representation

The private event plan specifies the intended content but is not treated as factual ground truth. Instead, only information realized in the generated dialogue or verified from the corresponding visual observation is retained as an *observed fact*. These facts are organized into a temporal ledger that records their provenance and validity over time. Each fact is associated with textual or visual evidence, its first valid session, and, when applicable, relations indicating that it has been superseded, retracted, or contradicted by a later observation.

This representation distinguishes factual evolution from contradiction: a fact that is valid at an earlier time remains part of the historical memory even when it is later updated. The ledger therefore supports queries referring to different temporal cutoffs and prevents information introduced in later sessions from being used as evidence for an earlier memory state. Each fact is linked back to the event from which it was observed, enabling subsequent QA annotations to remain fully traceable.

D.3 QA and Evidence Construction

QA specifications are generated from the finalized observed memory rather than from the private event plan. MemVista covers nine memory-oriented categories: factual retrieval (FR), visual search (VS), test-time learning (TTL), temporal reasoning (TR), visual reasoning (VR), multi-hop reasoning (MR), knowledge revision (KR), conflict detection (CD), and abstention or absence (AR). Each specification defines a reference answer, the relevant temporal cutoff, required factual or visual information, and one or more supporting events. The structured specification is then converted into a natural-language question.

Supporting evidence is annotated at the event level using provenance from the observed-fact ledger. Each QA is associated with one or more sufficient evidence sets, including required visual evidence when applicable. Alternative evidence sets are retained when multiple event combinations can support the same answer, while distractor events or images are recorded as hard negatives.

D.4 Quality Control and Data Filtering

Data quality is controlled throughout both multimodal generation and QA construction. Generated events are checked for consistency across dialogue and visual content, while QA candidates are validated against the observed memory to ensure that the answer and supporting evidence are temporally and semantically well grounded.

A further two-stage filtering procedure is applied to ensure that each QA genuinely depends on the annotated memory. In the first stage, the question is answered using its annotated supporting memory, and only instances for which the reference answer can be recovered correctly are retained. This removes ambiguous or intrinsically under-specified queries. In the second stage, the annotated supporting memory is removed while the remaining conversation history is preserved. The resulting answer must either be incorrect or explicitly abstain; otherwise, the instance is rejected. This stage filters questions affected by evidence leakage or answerable from model-internal knowledge without accessing the required memory. Only instances satisfying both conditions are retained in the final training corpus. For more challenging categories such as TTL and MR, automatically generated QAs are additionally reviewed by human annotators and reconstructed when necessary to ensure their validity and evidence grounding.

Table 7: Task composition of MemVista.

Category	Abbr.	#QA
Factual Retrieval	FR	178
Visual Search	VS	173
Test-Time Learning	TTL	162
Temporal Reasoning	TR	181
Visual Reasoning	VR	157
Multi-hop Reasoning	MR	148
Knowledge Revision	KR	148
Conflict Detection	CD	183
Abstention / Absence	AR	176
Total		1,506

Table 8: Statistics of the MemVista dataset.

Statistic	Count
Long-term scenarios	21
Conversational sessions	278
Dialogue events	3,646
Historical images	918
Total QA instances	1,506
Queries with an input image	269

Table 9: Additional qualitative case studies on Mem-Gallery. The table reports the original tool-call trajectories produced by the model for two successful examples and one representative failure, together with brief analysis of the behavior.

Query	Tool-Call Trajectory	Analysis
How many Collared Luxe Faux Fur Jackets did Chloe view?	RETRIEVE `{"method":"hybrid", "query":"Chloe Collared Luxe Faux Fur Jackets viewed"}` → INSPECT_IMAGE → STOP	**Correct**. Hybrid retrieval first locates candidate memories related to Chloe and the Collared Luxe Faux Fur Jackets. INSPECTIMAGE then verifies the retrieved visual evidence and supports counting the relevant jackets before termination.

When did Lumi first visit Lena's brother's house?	RETRIEVE
	`{"method":"bm25", "query":"Lumi Lena brother house visit date"}`
	→ RETRIEVE
	`{"method":"bm25", "query":"Lumi first visit Lena brother house"}`
	→ SEARCH_METADATA
	`{"field":"timestamp", "op":"contains", "value":"2024-09-10"}`
	→ STOP
		Correct. The retrieved turn is dated September 10, 2024, but states that the first visit occurred "the day before yesterday." Temporal reasoning over the evidence yields the correct date, September 8, 2024.

Please search for all images representing the dog breed that Finn represents.	RETRIEVE
	`{"method":"hybrid", "query":"dog breed Finn"}`
	→ INSPECT_IMAGE → RETRIEVE
	`{"method":"hybrid", "query":"Welsh Springer Spaniel images"}`
	→ *repeated no-gain retrieval*
		Incorrect. Inspecting Daisy's image causes the policy to confuse her breed with Finn's. It then repeats the incorrect search and returns `D5:IMG_001` instead of Finn's three reference images.

D.5 DATASET STATISTICS

The current MemVista release contains 21 long-term scenarios comprising 278 conversational sessions, 3,646 dialogue events, and 918 historical images. A total of 1,506 QA instances are included, covering nine memory-oriented task categories, with 269 queries accompanied by an input image. The overall dataset statistics and task composition are summarized in Tables 8 and 7, respectively.

E ADDITIONAL CASE STUDIES

To complement the qualitative examples in the main paper, Table 9 presents two additional successful retrieval trajectories and one representative failure case from Mem-Gallery. These examples cover visually grounded counting, temporal grounding, and entity confusion, illustrating both the adaptive behavior learned by the memory access policy and its remaining limitations.

The first two cases illustrate complementary forms of adaptive memory access. For the jacket-counting query, hybrid retrieval identifies relevant visual candidates and INSPECTIMAGE is then used to verify the target items before counting. For the temporal query, BM25 retrieval locates the relevant visit and metadata search grounds the supporting turn, allowing the relative expression "the day before yesterday" to be resolved to September 8, 2024. Together, these examples show how different tool combinations can support visually grounded and temporally grounded queries.

A representative failure asks for all images corresponding to the dog breed represented by Finn. The source memory identifies Finn as an Irish Water Spaniel, while retrieved evidence also contains another dog, Daisy, identified as a Welsh Springer Spaniel. After inspecting Daisy's image, the policy shifts its search to "Welsh Springer Spaniel images" and subsequently repeats the same retrieval request. The trajectory eventually returns Daisy's image, `D5:IMG_001`, instead of the three reference images associated with Finn's breed. This case suggests that entity consistency can still be challenging when retrieval is influenced by a competing semantic association.

F IMPLEMENTATION DETAILS

Models and infrastructure. Qwen3.5-4B is used as the memory access policy for all trainable variants. In the main OPSD experiment, a frozen Qwen3.5-4B counterpart serves as the privileged teacher, while Qwen3.5-9B is used for auxiliary inference throughout training and evaluation. These roles use separate prompts and are not jointly optimized. All training experiments are conducted on NVIDIA A100 GPUs.

State-level OPSD training. The initial learning rate is 5×10^{-7} with cosine decay to 1×10^{-7}, no warmup, and weight decay 0.01. The global query batch size and mini-batch size are both 24, with one on-policy rollout sampled per query. The maximum prompt and response lengths are 16,384 and 2,048 tokens, respectively. Training rollouts use temperature 1.0, top-p 1.0, and unrestricted top-k sampling. Training is configured for three epochs. Both entropy regularization and additional reference KL are disabled.

GRPO training. GRPO uses the same Qwen3.5-4B initialization, retrieval backend, tool interface, and interaction budget. Each query is sampled four times to form a GRPO group, resulting in 96 trajectories for a query batch size of 24. Student training uses sequence parallel size 4 and rollout tensor parallel size 1. The learning rate is fixed at 5×10^{-7} with no warmup, and weight decay is set to 0.01. Rollouts use temperature 0.8, top-p 1.0, and unrestricted top-k sampling. Both the maximum prompt and response lengths are 16,384 tokens. The PPO clipping range is 0.2, group advantages are normalized with $\epsilon = 10^{-6}$.

The GRPO reward is based on the mapped LLM judge score with additional penalties for repeated actions and unsuccessful termination:

$$R = S - \min\left(0.3,\ 0.02N_{\text{repeat}} + P_{\text{terminal}}\right), \tag{11}$$

where $S = m(s_i) \in \{0, 0.5, 1\}$ is the mapped judge score defined in Eq. (14). N_{repeat} counts repeated actions with identical tool names and arguments. P_{terminal} denotes the termination penalty, set to 0.1 when any unsuccessful termination condition is triggered. The total shaping penalty is capped at 0.3. For the i-th trajectory in a group of $G = 4$,

$$A_i = \frac{R_i - \text{mean}_{j=1}^{G} R_j}{\text{std}_{j=1}^{G}(R_j) + 10^{-6}}, \tag{12}$$

and the resulting advantage is applied to valid response tokens through the clipped policy objective.

Agent execution and evaluation. Each query is allowed at most 10 memory access actions including STOP. Only one tool is executed at each decision step.

Evaluation is conducted on 1,711 Mem-Gallery questions and 371 open-ended questions from the eight core MemEye scenarios. The Qwen3.5-4B policy is responsible only for memory access. After retrieval terminates, a separate answer model receives the query, optional query image, and final selected evidence to generate the response. Two answer-and-judge configurations are reported. The first uses Qwen3.5-9B for both answer generation and evaluation, while the second uses GPT-5.5 for both components. External baselines retain their released memory access mechanisms and recommended configurations, while final answer generation and evaluation are standardized under the corresponding answer-and-judge setting.

Evaluation metrics. The LLM judge produces a raw score

$$s_i \in \{0, 0.25, 0.5, 0.75, 1\}, \tag{13}$$

which is mapped to

$$m(s_i) = \begin{cases} 0, & s_i < 0.25, \\ 0.5, & 0.25 \le s_i < 0.75, \\ 1, & s_i \ge 0.75. \end{cases} \tag{14}$$

The reported Judge score is the mean mapped score,

$$\text{Judge} = \frac{1}{N} \sum_{i=1}^{N} m(s_i), \tag{15}$$

whereas strict accuracy counts only samples receiving the maximum raw score:

$$\text{Acc.} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[s_i = 1]. \tag{16}$$

Thus, a raw judge score of 0.75 contributes 1 to the mapped Judge metric but is not counted as strictly correct. Avg. Actions denotes the average number of memory access actions per query and includes STOP. Evidence recall is computed from the final evidence set against the annotated supporting events.

G ADDITIONAL EXPERIMENTS

G.1 EVIDENCE RETRIEVAL QUALITY

To directly evaluate evidence retrieval quality, we compare the final evidence set against the annotated supporting evidence on both Mem-Gallery and MemEye. Let E^* denote the reference evidence set and E_T the evidence retained in the final evidence set. We report four complementary metrics. *Full Recall* denotes the proportion of queries for which all reference evidence is retrieved, i.e., $E^* \subseteq E_T$. *Partial Recall* denotes cases where at least one reference item is retrieved but the evidence set remains incomplete, while *No Recall* denotes cases where none of the annotated supporting evidence is present in the final evidence set. We further report the average *Recall Ratio*,

$$\text{Recall} = \frac{|E_T \cap E^*|}{|E^*|}, \tag{17}$$

which measures the fraction of annotated supporting evidence recovered by the final evidence set.

Table 10 compares Base Agent, GRPO, and OPSD on both benchmarks. This evaluation separates retrieval quality from downstream answer generation and directly measures how policy optimization affects the completeness of acquired evidence.

Results. As shown in Table 10, both learned policies substantially improve evidence coverage over the Base Agent on the two benchmarks. On Mem-Gallery, OPSD increases Full Recall from 68.07% to 79.74% and Recall Ratio from 80.24% to 88.11%, while reducing No Recall from 8.92% to 4.39%

Table 10: Evidence retrieval quality and tool-use efficiency on Mem-Gallery and MemEye. Full Recall, Partial Recall, and No Recall denote complete, incomplete, and zero supporting-evidence retrieval, respectively. Recall Ratio is the average fraction of reference evidence recovered. Avg. Calls denotes the average number of memory-access actions per query.

Method	Full Recall (%) ↑	Partial Recall (%)	No Recall (%) ↓	Recall Ratio (%) ↑	Avg. Calls ↓
Mem-Gallery					
Base Agent	68.07	23.02	8.92	80.24	6.45
GRPO	78.49	16.23	5.28	87.95	6.03
OPSD	**79.74**	15.87	**4.39**	**88.11**	**2.89**
MemEye					
Base Agent	48.65	35.41	15.95	66.65	7.81
GRPO	54.41	33.16	12.43	71.07	6.58
OPSD	**55.14**	32.97	**11.89**	**71.66**	**5.66**

Table 11: Distribution of memory access tool calls on Mem-Gallery and MemEye. Values denote the percentage of each operation among all non-STOP actions.

Method	Retrieve (%)	SearchMetadata (%)	ExpandContext (%)	InspectImage (%)
Mem-Gallery				
Base Agent	86.22	2.04	2.09	9.66
GRPO	56.76	0.74	27.90	14.59
OPSD	88.56	5.75	0.41	5.28
MemEye				
Base Agent	76.14	3.32	11.27	9.27
GRPO	56.37	0.28	27.88	15.47
OPSD	74.49	12.44	0.11	12.96

and the average number of memory-access actions from 6.45 to 2.89. Similar improvements are observed on MemEye, where Full Recall rises from 48.65% to 55.14% and No Recall decreases from 15.95% to 11.89%. OPSD also requires fewer actions on MemEye, averaging 5.66 calls compared with 7.81 for the Base Agent and 6.58 for GRPO. OPSD consistently improves over GRPO in evidence coverage, achieving higher Full Recall and Recall Ratio together with lower No Recall on both benchmarks. The reduction in Partial Recall is accompanied by an increase in Full Recall, indicating that more queries progress from incomplete evidence coverage to complete retrieval. These results provide direct evidence that state-level OPSD improves the ability of the memory access policy to acquire sufficient supporting evidence while using a more efficient retrieval trajectory.

G.2 TOOL USAGE DISTRIBUTION

To further examine how policy optimization changes memory access behavior, we analyze the distribution of tool calls on Mem-Gallery and MemEye. The four retrieval modes are grouped under RETRIEVE, while SEARCHMETADATA, EXPANDCONTEXT, and INSPECTIMAGE are reported separately. STOP is excluded from the statistics, and the remaining tool calls are normalized to obtain the relative usage frequency of each operation. Table 11 compares Base Agent, GRPO, and OPSD under the same tool interface.

Table 11 reveals distinct tool-use patterns across the three policies. Most notably, OPSD exhibits a clear shift in tool composition between the two benchmarks. On Mem-Gallery, RETRIEVE accounts for 88.56% of non-stop actions, whereas on MemEye its proportion decreases to 74.49%, accompanied by substantial increases in SEARCHMETADATA from 5.75% to 12.44% and INSPECTIMAGE from 5.28% to 12.96%. This shift is consistent with the characteristics of MemEye, which places greater emphasis on fine-grained visual evidence and frequently requires more precise localization of relevant events before visual verification. The increased use of metadata search and image in-

spection therefore suggests that OPSD adapts its tool choices to the information requirements of the query rather than relying on a fixed retrieval pattern.

In contrast, GRPO produces nearly identical tool distributions on Mem-Gallery and MemEye, with approximately 56% of actions assigned to RETRIEVE and 28% to EXPANDCONTEXT on both benchmarks. The Base Agent similarly shows only limited variation across the two datasets. Together with the evidence retrieval results above, these observations suggest that state-level supervision enables more query-dependent composition of memory access operations, whereas the Base and GRPO policies exhibit less differentiated tool-use behavior across tasks.

H LIMITATIONS

This work focuses on long-term multimodal memory involving interleaved text and images, and does not yet consider richer modalities such as audio or video, which may require different retrieval and evidence inspection mechanisms. In addition, the current evaluation is conducted on Mem-Gallery and MemEye, while MemVista is synthetically constructed; broader validation on real-world multimodal interaction histories remains an important direction for future work.

I CORE SYSTEM PROMPTS

We provide the two prompts central to state-level OPSD: the state verifier used to construct privileged information and the privileged teacher used for action supervision. Dynamic fields are replaced by placeholders. Other prompts used for MemVista construction, tool-side auxiliary components, answer generation, and evaluation will be released together with the code.

I.1 STATE VERIFIER PROMPT

The verifier receives the current public evidence state together with the reference answer as a private rubric and produces a structured diagnosis of evidence sufficiency and the remaining information gap.

```
You are the OPD-MM state verifier. Decide whether the public state entails the answer required
 by the private rubric. Judge meaning rather than exact wording or formatting.

Question: {{QUESTION}}
Public state: {{PUBLIC_OBSERVATION_JSON}}
Private rubric -- use silently and never quote or describe: {{GOLD_ANSWER}}

Apply these rules:
1. Every part of a positive or multi-part answer needs public support. A list/count needs a
complete set; examples or an ordinal gap remain incomplete, and a "last" item does not fill a
missing ordinal.
2. A question asking whether a claim conflicts with memory is always a conflict task.
Compatible evidence supports no conflict; incompatible information for the same entity or
event supports conflict. Absence proves neither.
3. Paired dialogue and images are one event. Image identification requires only a public
image_id and matching description; inspect raw visual detail only when the queried visual
attribute is undescribed or ambiguous.
4. Treat the rubric as absence only when it means absent, unknown, or not mentioned. Absence
is sufficient only after multiple complementary query-grounded searches cover the required
modalities. Search progress never proves a positive fact or conflict.

Choose one missing type: none; no_public_evidence; irrelevant_evidence;
missing_metadata_constraint; candidate_set_too_broad; missing_neighbor_context;
missing_temporal_order; missing_raw_visual_detail; incomplete_coverage;
missing_factual_support.

Return JSON only:
{
  "evidence_sufficient": boolean,
  "missing_evidence_type": "none|no_public_evidence|irrelevant_evidence|
missing_metadata_constraint|candidate_set_too_broad|missing_neighbor_context|
missing_temporal_order|missing_raw_visual_detail|incomplete_coverage|missing_factual_support",
  "reason": "at most 35 words identifying the relevant public support and unresolved query
requirement"
}
```

I.2 PRIVILEGED TEACHER PROMPT

The privileged teacher receives the sanitized verifier diagnosis in addition to the same public state available at the corresponding student decision point. The reference answer itself is not included.

```
You are the OPD-MM action teacher for one student-visible state. Produce exactly one tool
action.

Question: {{QUESTION}}
Verifier diagnostic: {{VERIFIER_FEEDBACK_JSON}}
Public action history: {{ACTION_HISTORY_JSON}}
Current public observation: {{PUBLIC_OBSERVATION_JSON}}

Apply this gate before considering the question modality or any tool description:
- evidence_sufficient=true and no observation error: output STOP immediately. Every non-STOP
tool is invalid.
- evidence_sufficient=false: choose one non-STOP tool that addresses the diagnosed gap.

You are not answering the question. Verifier feedback is a private diagnostic, not evidence
or an action command. All arguments must be derivable from the question, public history, and
public observation. Do not copy verifier.reason or private content into an argument. A
RETRIEVE.query may use only public question/history/observation text.

Choose the schema-described tool whose effect addresses the gap and whose preconditions hold.
Do not repeat an identical action when the latest observation is unchanged, and never invent
an argument not supported by public state.

Emit exactly one function call with no reasoning, markdown, JSON, hidden memory IDs, or
unknown arguments. The chat template supplies the tool descriptions and serialization.

Output the corrected action now. Begin with <tool_call> and include nothing else.
```