MNEME: AN ODYSSEY THROUGH RECONSTRUCTIVE MEMORY FOR PERSONAL DEEP IMAGE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Image retrieval enables users to find relevant photos in large image collections, such as their photo albums. Common embedding-based methods rank images independently by query–image similarity, but personal album queries may identify targets through contextual clues found in other photos. An effective search must therefore recover cross-photo relations and use them to locate and verify target images. To this end, we introduce **MNEME**, a memory-augmented agent for reconstructive search over personal photo collections. Rather than using memory only to preserve or reuse agent search experience, MNEME reconstructs the visual history itself as a persistent relational memory. Its Album Memory Graph links grounded photo evidence, recurring visual instances, and overlapping episodic contexts, while Self-Evolving Search Skills induce and iteratively refine reusable search guidance from same-query success–failure trajectory contrasts on auxiliary tasks. At inference time, a dual-route strategy combines memory-guided reconstruction with memory-independent evidence search, and reconciles their candidates against raw images and metadata. Experiments on DISBench and CamRoll show that MNEME consistently outperforms strong retrieval and memory-augmented baselines across both tested backbones. Ablations further indicate that persistent album memory alone is insufficient; its benefits emerge when combined with reliable short-term memory, learned memory search guidance, and complementary memory-guided and direct search.

1 INTRODUCTION

Image retrieval enables users to find relevant visual content in large image collections. Embedding-based methods commonly encode images and natural-language queries in a shared space and rank each image independently by query–image similarity (Radford et al., 2021; Li et al., 2026; Zhou et al., 2025). Personal photo albums, however, record connected experiences, with people, places, and activities recurring across photos (Rodden & Wood, 2003; Frohlich et al., 2002). Users may describe a desired photo through an associated event or companion, including clues visible only in other images. Resolving such queries requires connecting evidence across photos to locate and verify the targets.

For example, a user may search for a photo containing metal chairs taken on the day they saw a wall illustrated with fresh ingredients (Figure 1). The wall identifies the day, whereas the chairs specify the target content; these clues need not co-occur. Retrieving chair photos alone may return images from unrelated days, so the system must locate the wall, establish the temporal relation, and verify candidate targets. Deep image search captures this dependence on contextual evidence distributed across a visual history (Deng et al., 2026). Related benchmarks also examine personal intent and camera-roll reasoning beyond direct visual matching (Xu et al., 2026; Nguyen et al., 2026a).

Tool-using agents address this setting through semantic retrieval, metadata filtering, and visual inspection (Yao et al., 2023; Deng et al., 2026). Memory-based methods further support album navigation and reuse of search experience (Nguyen et al., 2026a; Lyu et al., 2026; Shinn et al., 2023; Zhao et al., 2024; Ouyang et al., 2025). However, making contextual evidence accessible does not determine which relations to follow, when to revise a search scope, or which candidates require inspection. We therefore investigate how to couple navigable album relations with memory-graph search guidance while grounding final decisions in raw images and metadata.

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.

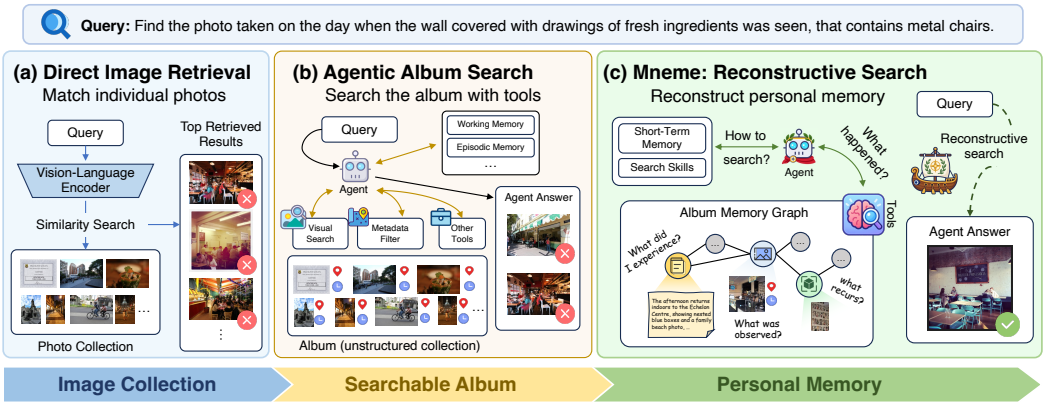


Figure 1: Comparison of personal image search paradigms. (a) Direct retrieval matches individual photos, (b) agentic search explores an unstructured album, while (c) MNEME recollects cross-photo context from reconstructed album memory to locate and verify target images.

To this end, we introduce **MNEME**, a memory-augmented agent that couples structured album memory with memory-graph search skills for personal deep image search. We formulate this process as *reconstructive search*: recovering latent query context from distributed album evidence, using the recovered context to identify search scopes, and verifying target photos against visual and metadata evidence. MNEME reconstructs each album as an Album Memory Graph with three types of nodes: photo nodes retain the original images and their metadata; clue nodes represent recurring people and objects in photos; and story nodes summarize temporal, spatial, and episodic memories. Edges link clues and stories to their supporting photos, allowing the agent to follow cross-photo associations, identify relevant search scopes, and return to the underlying evidence for verification.

To guide memory use, MNEME induces *Self-Evolving Search Skills* from success–failure contrasts on synthetic skill-induction sets constructed from auxiliary albums disjoint from the evaluation users. Multiple rollouts for the same query provide contrasting trajectories that extract guidance on scope selection, relation traversal, verification, and failure recovery. Within- and cross-generation comparisons iteratively revise this guidance, and the final skill bank is frozen before evaluation without updating the backbone parameters.

At inference time, MNEME runs two complementary routes: a *recollective route* with access to the Album Memory Graph and a *direct evidence route* using standard retrieval tools. Each maintains its own Evidence-Aware Short-Term Memory, and a final verifier reconciles candidates from both routes against raw images and metadata. This design separates the memory-guided search from the final evidence-based verification.

Extensive experiments on DISBench (Deng et al., 2026) and CamRoll (Nguyen et al., 2026a) demonstrate the effectiveness of MNEME across image-set retrieval and camera-roll question answering. In DISBench, MNEME consistently outperforms baselines in both intra-event and inter-event retrieval. On CamRoll, our approach also enhances evidence recall and improves response quality, thereby supporting its applicability beyond image-set retrieval tasks. Progressive ablations show that richer memory alone does not guarantee better search; its effectiveness depends on how reliably the agent can use and verify recalled context.

To summarize, our contributions are threefold: (1) We formulate personal deep image search as *reconstructive search* and introduce an **Album Memory Graph** that represents the album as persistent relational memory over grounded photos, recurring visual instances, and overlapping episodic contexts. (2) We introduce **Self-Evolving Search Skills** that learn how to navigate album memory from same-query success–failure and cross-generation trajectory contrasts, yielding reusable guidance for scope selection, relation traversal, verification, and recovery. (3) We develop a **Dual-Route** inference framework that combines memory-guided reconstruction with memory-independent evidence retrieval and reconciles their candidates against raw images and metadata to mitigate errors from misleading memory.

2 Related Work

Multimodal Agentic Search Recent multimodal retrieval has evolved from embedding-based cross-modal matching methods (Radford et al., 2021; Li et al., 2026; Faysse et al., 2025; Zhang et al., 2026) toward agentic search, where multimodal models iteratively reason, retrieve, and invoke tools to gather evidence. MMSearch-R1 (Wu et al., 2026), WebWatcher (Geng et al., 2025), and DeepMMSearch-R1 (Narayan et al., 2025) enable MLLMs to actively issue multimodal queries and refine search strategies based on intermediate observations. More closely related to our setting, ImageSeeker (Deng et al., 2026), PhotoBench (Xu et al., 2026), and CamRoll-Agent (Nguyen et al., 2026a) study agentic reasoning over long-term personal visual histories, where relevant evidence may be distributed across multiple photos. While these works establish the importance of active exploration beyond query–image similarity, MNEME focuses on reconstructive search over persistent visual memory, explicitly combining structured cross-photo album memory with transferable search experience.

Agent Memory Retrieval from contextual or partial cues has long been studied in episodic and associative memory (Tulving & Thomson, 1973; Marr et al., 1991). This perspective motivates our use of cross-photo associations to recover relevant album context. Agent memory has evolved from raw interaction histories toward structured and self-evolving mechanisms that support long-horizon reasoning and experience reuse. Text-centric approaches explore reusable workflows and reasoning strategies (Wang et al., 2025; Ouyang et al., 2025), dynamically organized memory graphs (Xu et al., 2025), scalable memory extraction and lifecycle management (Chhikara et al., 2025; Li et al., 2025), and transferable skill guidance distilled from agent trajectories (Fang et al., 2026). Recent multimodal systems further extend persistent memory beyond text: M3-Agent (Long et al., 2025) builds entity-centric episodic and semantic memories from visual and auditory streams, while MIRIX (Wang & Chen, 2025), WorldMM (Yeo et al., 2025), and MemVerse (Liu et al., 2025) organize heterogeneous multimodal experiences for long-horizon reasoning. VisualMem (Nguyen et al., 2026b) further emphasizes preserving personalized visual evidence rather than reducing images to generic textual memories. Most closely related, PhotoCraft (Lyu et al., 2026) equips deep-image-search agents with hierarchical working, episodic, and semantic memory. In contrast, MNEME explicitly separates memory of the visual history from memory of the retrieval process: the Album Memory Graph organizes grounded cross-photo content, while Self-Evolving Search Skills distill transferable search and verification procedures.

3 Problem Formulation

Following DeepImageSearch (Deng et al., 2026), we study context-dependent image-set retrieval over a personal album $\mathcal{A} = \{a_i\}_{i=1}^{N}$. Each photo $a_i = \langle v_i, t_i, g_i \rangle$ contains visual content v_i and, when available, a capture time t_i and geographic location g_i. Given a natural-language query Q, the system returns a predicted set $\hat{\mathcal{R}} \subseteq \mathcal{A}$:

$$F : (\mathcal{A}, Q) \to \hat{\mathcal{R}}. \tag{1}$$

The goal is to recover all photos that satisfy the query, including conditions whose supporting evidence occurs elsewhere in the album.

We model the search as an iterative interaction between a vision-language model (VLM) agent and the album, following ReAct (Yao et al., 2023). The VLM serves as the agent backbone, reasoning over the natural-language query, visual evidence, metadata, and tool observations to decide subsequent actions. At step t, the agent selects an action u_t from the query and preceding interaction history, then receives a tool observation:

$$u_t \sim \pi_\theta(u \mid Q, \tau_{<t}), \qquad \tau_{<t} = (u_1, o_1, \ldots, u_{t-1}, o_{t-1}), \tag{2}$$

where π_θ denotes the VLM-parameterized agent policy. Actions retrieve candidate photos, filter metadata, inspect images, or access cross-photo relations. After T steps, the agent predicts

$$\hat{\mathcal{R}} = G_\theta(Q, \tau_{1:T}) \subseteq \mathcal{A}. \tag{3}$$

Rather than treating the album as an unstructured collection, MNEME reconstructs it as an Album Memory Graph $\mathcal{M}_A$, augmenting the agent's interaction space with graph-grounded operations for

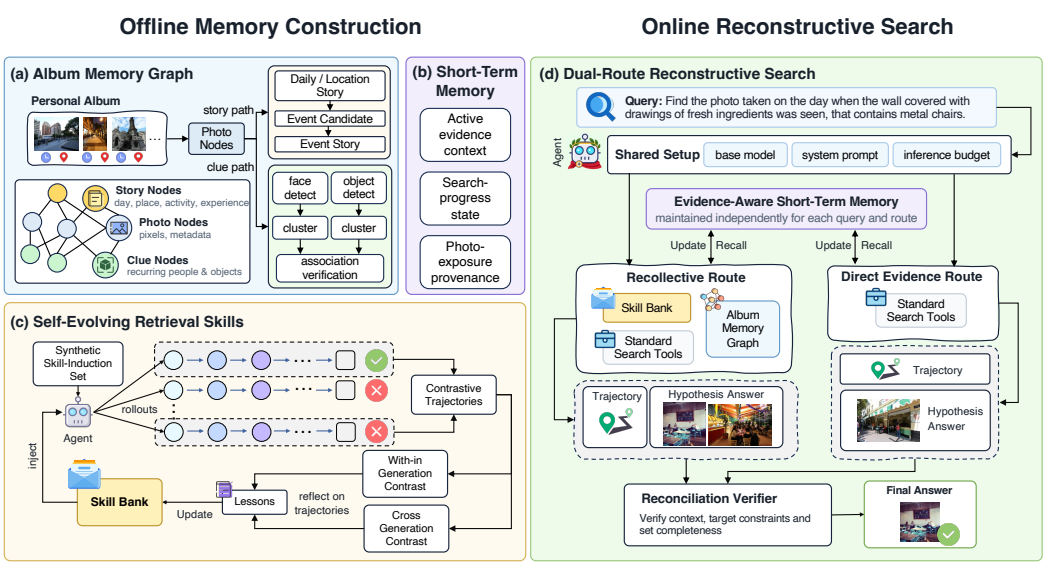


Figure 2: Overview of MNEME. Offline, we construct an Album Memory Graph from each album and induce search skills from contrasting trajectories on synthetic data. Online, memory-guided and direct routes search in parallel with Short-Term Memory, and their candidates are reconciled against raw images and metadata.

retrieving contextual memories, traversing cross-photo relations, and constructing evidence scopes, while preserving access to the underlying images and metadata for verification.

4 METHODOLOGY

Overview. MNEME performs *reconstructive search* by linking contextual cues across a personal album and checking candidate photos against the query (Figure 2). Its four components separate album structure, query-local evidence, reusable search guidance, and final answer verification:

- **Album Memory Graph (§4.1).** We organize each album into grounded photo memories, recurring visual clues, and episodic stories to make cross-photo context accessible.
- **Evidence-Aware Short-Term Memory (§4.2).** We maintain a bounded query-local state that preserves search progress, photo provenance, and evidence judgments across long trajectories.
- **Self-Evolving Search Skills (§4.3).** We distill guidance for scope selection, relation traversal, verification, and recovery from contrasting successful and unsuccessful trajectories on auxiliary albums.
- **Dual-Route Reconstructive Search (§4.4).** We run searches with and without album-memory access, then reconcile their results.

4.1 ALBUM MEMORY GRAPH

For each album, we construct an Album Memory Graph $\mathcal{M}_A = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes comprised of photo nodes $\mathcal{V}_P$, visual-clue nodes $\mathcal{V}_C$, and story nodes $\mathcal{V}_S$:

$$\mathcal{V} = \mathcal{V}_P \cup \mathcal{V}_C \cup \mathcal{V}_S. \tag{4}$$

Clue nodes $\mathcal{V}_C$ and story nodes $\mathcal{V}_S$ are linked to their supporting photos using edges $\mathcal{E}$, exposing recurring entities and temporal, spatial, or episodic associations (Figure 2 (a)). Note that a photo may participate in several stories and contain multiple recurring clues. The graph allows the agent to enter an album through different contextual references while retaining access to the underlying photo evidence.

Photo nodes $\mathcal{V}_P$. Each photo node represents the raw evidence of a photo in the album, including its image v_i and metadata (t_i, g_i). Photo nodes form the grounded basis of the memory graph, anchoring higher-level memories while preserving the original evidence for album-level reasoning and verification.

Clue nodes $\mathcal{V}_C$. We detect salient face and object regions, group visually consistent instances, and verify proposed cross-photo correspondences against direct visual evidence (see Appendix A for details). Each verified group forms a clue node linked to its supporting photos, enabling the agent to trace recurring people or objects across contexts. Correspondences are verified during graph construction, while query-time inspection determines whether linked photos satisfy the current query.

Story nodes $\mathcal{V}_S$. We use a Vision-Language Model to progressively consolidate photos and metadata into episodic memories at multiple scales. Photo-level descriptions are organized into daily, location, and activity memories through spatiotemporal grouping and event-coherence assessment, then synthesized into broader experience memories. Each story retains links to its supporting photos and, where applicable, constituent stories. These overlapping contexts enable coarse-to-fine search from relevant episodes to individual photos.

We build modality-specific indices over the three node types. Retrieved nodes provide entry points into photo scopes that the agent refines through metadata filtering and image inspection. For the query in Figure 1, the agent can locate a wall photo, follow its story links to the relevant day, and search that day's photos for chairs. The graph thus connects a contextual cue to a searchable photo set.

Graph access. The tools expose three capabilities. Context lookup retrieves story synopses or recurring visual clues. Scoped exploration searches a story's supporting photos, inspects photo-to-story and photo-to-clue connections, and applies time, location, and set constraints. Photo-evidence retrieval combines dense image and story-memory search, either across the album or within an explicit scope, and can inspect same-day neighborhoods around a candidate.

Story search can surface candidates outside the selected story so that the agent can revise its scope. Story synopses expose no Photo IDs, and temporal neighbors serve as proposed anchors. Pagination, scope inventories, and set certificates track coverage within the explored scope. The seven tool interfaces are described in Appendix B.

4.2 Evidence-Aware Short-Term Memory

The Album Memory Graph stores persistent album context; short-term memory retains the evidence and progress of the current search. To support long trajectories within limited context and image budgets, we maintain a bounded working state at step t:

$$\mathcal{M}_t = (\mathcal{E}_t, \mathcal{P}_t, \mathcal{X}_t). \tag{5}$$

Its components separate the evidence used for reasoning from the records needed to continue the search:

- **Active evidence context $\mathcal{E}_t$.** The query, recent tool observations, rendered photos, metadata, and intermediate judgments about candidate relevance.
- **Search-progress state $\mathcal{P}_t$.** Named scopes, retrieval arguments, pagination offsets, metadata filters, repeated requests, and tool counts.
- **Photo-exposure state $\mathcal{X}_t$.** Successfully loaded and presented Photo IDs, together with their source tools, presentation counts, most recent presentation turns, and image detail levels.

When the context or image budget approaches its limit, older messages in $\mathcal{E}_t$ are replaced by a bounded textual checkpoint. The checkpoint preserves the query goal, anchor–target distinction, supported and rejected observations, completed searches, named scopes, and next unresolved action. The progress and exposure states remain separate, so explored scopes and rendered Photo IDs survive history compression without being conflated with relevance judgments.

The state is reset for each query, and each inference route maintains its own instance (§4.4). If a route exhausts its tool-turn budget, an answer-only finalizer uses the retained state without further retrieval or image inspection.

4.3 Self-Evolving Search Skills

The Album Memory Graph provides contextual evidence, and short-term memory preserves the current search state. Neither, however, specifies which scope to explore, which relation to follow, or when to revise an unsuccessful search. Building on learning from agent experience, we address these decisions with a textual bank of *Self-Evolving Search Skills*, $\mathcal{S}^{(r)}$, at generation r. We induce the skills on synthetic skill-induction sets constructed from auxiliary YFCC-100M albums (Thomee et al., 2016), disjoint from the evaluation benchmarks in users, photos, questions, answers, and trajectories (Appendix C).

Starting with the Album Memory Graph and Short-Term Memory but no additional skills, we collect n rollouts per query in the skill-induction set. For image-set retrieval, a trajectory τ is successful when its predicted set exactly matches the reference set:

$$y(Q, \tau) = \mathbb{I}[\mathrm{EM}(Q, \tau) = 1]. \tag{6}$$

For queries with both successful and unsuccessful trajectories, a reflection model compares the runs under the same query and agent configuration. It extracts guidance about retrieval order, scope selection, relation traversal, verification, stopping, and recovery, then consolidates these lessons into the skill bank.

At each subsequent generation, we add the current skill bank to the agent's instructions and collect n new rollouts per query in the skill-induction set. Within-generation comparisons contrast successes and failures under the same guidance; cross-generation comparisons examine improvements and regressions on the same query as the guidance changes.

Let $\mathcal{L}^{(r)}$ denote the lessons extracted from these comparisons. The existing skill bank is revised as

$$\mathcal{S}^{(r+1)} = \mathrm{Revise}(\mathcal{S}^{(r)}, \mathcal{L}^{(r)}). \tag{7}$$

We run k generations and freeze $\mathcal{S}^{(k)}$ before evaluation. At inference time, the frozen bank is appended to the fixed system instructions. Skill induction updates this textual guidance, not the backbone parameters, and uses no evaluation labels.

4.4 Dual-Route Reconstructive Search

Search routes. Even with search guidance, a recalled story or association can lead the agent toward an irrelevant scope. MNEME therefore combines a *reconstructive route* with album-memory access and a *direct evidence route* without it. Each maintains the working state defined in §4.2; their proposed answers and retained evidence are reconciled after both searches finish.

Given a query Q, the two routes run in parallel:

$$(\mathcal{R}_d, \tau_d) = f_{\mathrm{direct}}(Q), \qquad (\mathcal{R}_r, \tau_r) = f_{\mathrm{reconstruct}}(Q; \mathcal{M}_A). \tag{8}$$

Here, $\mathcal{R}_d$ and $\mathcal{R}_r$ are the proposed answer sets, and τ_d and τ_r are the interaction traces. The direct evidence route uses standard retrieval tools, while the reconstructive route additionally accesses the Album Memory Graph for contextual hypotheses and cross-photo associations. Both routes use the same backbone and base system instructions, with the same search-budget limit per route. The total query cost includes both searches and final reconciliation.

Candidate pool. The routes may recover complementary evidence: one can locate plausible targets while the other resolves a contextual relation. To preserve useful candidates even when a final proposal is incomplete, we pool both proposed answer sets and selected intermediate candidates:

$$\mathcal{C} = \mathcal{R}_d \cup \mathcal{R}_r \cup R_K(\tau_d) \cup R_K(\tau_r), \tag{9}$$

where $R_K(\tau)$ selects a small number of recently inspected photos from trace τ. Including these photos preserves candidates that a route inspected but omitted from its final proposal. The pool is closed during reconciliation: this stage can select among the retained candidates but cannot discover targets outside the pool.

Table 1: Performance on DISBench and CamRoll. For each metric, **best** and second-best results are marked within each backbone, including ties. All metrics are percentages except Judge (0–10).

Method	DISBench						CamRoll			
	Intra-Event		Inter-Event		Overall		Multiple-Choice		Free-Form	
	EM	F1	EM	F1	EM	F1	Recall	Acc.	Recall	Judge
Qwen3-VL-32B-Instruct										
Naïve RAG	1.8	14.5	1.5	15.5	1.6	15.1	61.4	68.5	61.4	4.1
ImageSeeker	10.5	23.0	6.2	17.3	8.2	20.0	74.2	61.5	69.4	3.1
CamRoll-Agent	10.5	17.6	6.2	14.0	8.2	15.7	89.8	75.5	90.1	4.3
PhotoCraft	14.0	26.5	4.6	23.1	9.0	24.7	81.9	78.5	77.7	4.5
MNEME (Ours)	**17.5**	**30.7**	**10.8**	**29.5**	**14.0**	**30.0**	**90.3**	**79.0**	**93.8**	**5.6**
Qwen3.5-27B										
Naïve RAG	7.0	19.8	1.5	16.6	4.1	18.1	61.4	77.0	61.4	4.0
ImageSeeker	19.3	36.2	10.8	35.2	14.8	35.6	91.6	62.5	85.9	4.9
CamRoll-Agent	19.3	40.3	1.5	16.9	9.8	27.8	95.0	78.0	95.3	4.3
PhotoCraft	10.5	23.5	9.2	26.9	9.8	25.3	87.5	73.0	85.6	5.8
MNEME (Ours)	**24.6**	**46.0**	**21.5**	**38.2**	**23.0**	**41.8**	**95.5**	**82.0**	**95.7**	**6.2**

Evidence reconciliation. For each candidate $p \in \mathcal{C}$, the verifier receives the query, both proposed answer sets, compressed route traces, and the candidate's raw image and camera metadata. It assesses whether the retained evidence supports the contextual relation and whether the candidate satisfies the target predicates and exclusions. An anchor can help establish context without itself qualifying as an answer. The verifier also checks coverage within the candidate pool; this check does not establish completeness over the entire album. The final prediction is

$$\hat{\mathcal{R}} = V_\theta\left(Q, \mathcal{R}_d, \mathcal{R}_r, \widetilde{\tau}_d, \widetilde{\tau}_r, \{E(p)\}_{p\in\mathcal{C}}\right), \qquad \hat{\mathcal{R}} \subseteq \mathcal{C}, \tag{10}$$

where $E(p)$ contains the raw image and metadata of candidate p. Reconciliation selects the final set from the retained pool, performs no additional retrieval, and receives no reference answers or evaluation labels.

5 Experiments

5.1 Experimental Settings

We use eight NVIDIA A800-SXM4-80GB GPUs. Each search route has a maximum budget of 30 reasoning turns unless otherwise specified; total inference cost includes both parallel routes and candidate reconciliation. Skill induction uses four rollouts per query over three generations, after which the skill bank is frozen without updating backbone parameters.

Datasets and Metrics. We evaluate image-set retrieval on **DISBench** (Deng et al., 2026): 122 queries over 109,467 photos from 57 users, covering intra- and inter-event reasoning. We report exact match (EM) and F1 by query type and overall. For camera-roll question answering, we stratified sampled 200 **CamRoll** (Nguyen et al., 2026a) questions (40 semantic and 160 episodic) from 20 users, reporting evidence recall, multiple-choice accuracy, and free-form response quality scored by GPT-4o (OpenAI, 2024) on a 0–10 scale (Appendix E). Other metrics are percentages; higher is better.

Baselines. We compare MNEME with **Naïve RAG**, which predicts from similarity-ranked photos; **ImageSeeker** (Deng et al., 2026), which combines retrieval, metadata operations, and visual inspection; **CamRoll-Agent** (Nguyen et al., 2026a), which navigates hierarchical album memories; and **PhotoCraft** (Lyu et al., 2026), which combines working, episodic, and semantic memory. All methods are evaluated with Qwen3-VL-32B-Instruct (Team, 2025) and Qwen3.5-27B (Team, 2026), with comparisons made within each backbone.

Table 2: Cumulative ablation study on DISBench. Parentheses show percentage-point changes from the preceding row.

Configuration	EM	F1
Base Agent	14.8	35.6
+ Short-Term Memory	18.0 (+3.2)	37.0 (+1.4)
+ Album Memory Graph	14.8 (-3.2)	34.8 (-2.2)
+ Search Skills	19.7 (+4.9)	38.3 (+3.5)
+ Dual-Route Search	**23.0** (+3.3)	**41.8** (+3.5)

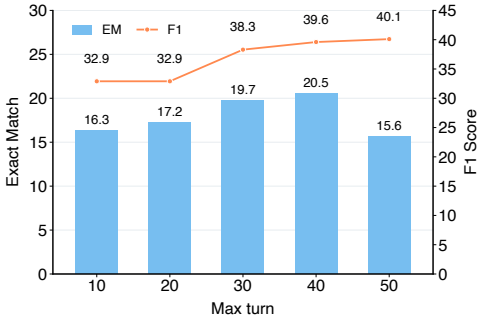


Figure 4: Performance versus reasoning-turn budget.

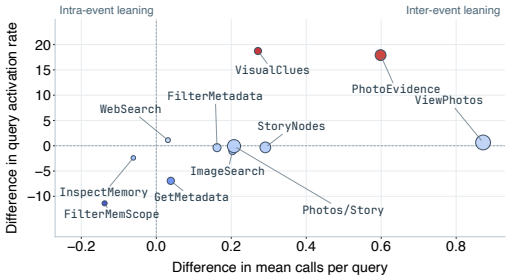


Figure 3: Differences in tool-call frequency and query activation.

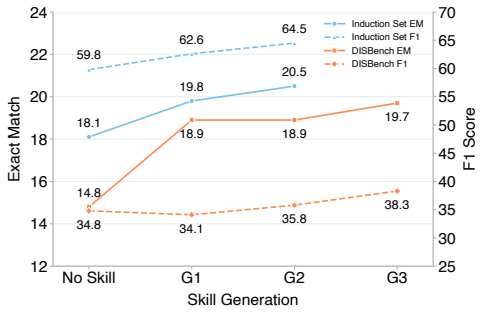


Figure 5: Search skill evolution on the skill-induction set (Induction) and DISBench.

5.2 Main Results

Table 1 summarizes the main results. We highlight three findings.

Overall Performance on DISBench. MNEME achieves the best overall EM and F1 under both backbones. With Qwen3-VL-32B-Instruct, it obtains 14.0 EM and 30.0 F1, outperforming PhotoCraft, the strongest baseline on both metrics, by 5.0 and 5.3 percentage points, respectively. With Qwen3.5-27B, MNEME reaches 23.0 EM and 41.8 F1, surpassing ImageSeeker by 8.2 and 6.2 points. These consistent improvements across backbones indicate gains in both exact answer-set recovery and partial answer overlap.

Intra- and Inter-Event Retrieval. MNEME further achieves the best performance on both intra- and inter-event queries under each backbone. With Qwen3.5-27B, it obtains 24.6 EM and 46.0 F1 on intra-event queries, improving EM by 5.3 points over the tied strongest baselines, ImageSeeker and CamRoll-Agent, and F1 by 5.7 points over CamRoll-Agent. On inter-event queries, MNEME reaches 21.5 EM and 38.2 F1, outperforming ImageSeeker by 10.7 and 3.0 points, respectively. With Qwen3-VL-32B-Instruct, MNEME achieves 17.5/30.7 EM/F1 on intra-event queries and 10.8/29.5 on inter-event queries, with EM improvements of 3.5 and 4.6 points over the strongest baselines, respectively. The larger EM improvements on inter-event queries suggest that MNEME is particularly effective at recovering complete answer sets when relevant evidence spans multiple events, although these results do not isolate the contribution of any individual component.

Camera-roll question answering. On the stratified sampled CamRoll subset, MNEME leads all four reported metrics under both backbones. With Qwen3-VL-32B-Instruct, it obtains 90.3% multiple-choice evidence recall and 79.0% accuracy, together with 93.8% free-form evidence recall and a Judge score of 5.6. These exceed the strongest baseline for each metric by 0.5, 0.5, and 3.7 percentage points, and 1.1 Judge points, respectively. With Qwen3.5-27B, the scores rise to 95.5%, 82.0%, 95.7%, and 6.2, with corresponding gains of 0.5, 4.0, and 0.4 percentage points, and 0.4 Judge points. Answer quality improves despite small recall gains, indicating that recall alone does not capture answering performance. These findings concern the sampled subset.

5.3 Ablation Studies

To examine the progression toward reconstructive search, we cumulatively introduce components on DISBench (Table 2). Each addition is evaluated within the preceding configuration, not independently.

Short-term memory improves both EM and F1, supporting query-local evidence retention. Adding the Album Memory Graph alone reverses these gains, indicating that access to a richer memory structure does not by itself ensure effective retrieval. The additional nodes, relations, and tools may complicate search decisions without suitable guidance. Introducing search skills then improves both metrics beyond the STM-only configuration, suggesting more effective memory use with learned guidance.

Dual-route search performs best, supporting complementary direct and memory-guided retrieval. This improvement reflects the full inference design, including reconciliation and additional computation; it cannot be attributed to reconciliation alone. Together, these results support coupling evidence retention, structured memory navigation, and complementary search.

5.4 Quantitative Analysis

Effect of Self-Evolution To assess skill induction and refinement, we track successive generations (Figure 5). On the skill-induction set, EM and F1 both rise. On DISBench, G1 sharply improves EM but slightly lowers F1; G2 improves F1 while EM remains stable, and G3 improves both. Higher induction-set F1 suggests easier partial recovery, but comparable EM and data reuse preclude broader difficulty claims. Initial EM gains support skill induction, and subsequent gains support refinement, with final improvements of 4.9 EM and 3.5 F1 points over no skills.

Effect of Maximum Reasoning Turns We vary the turn budget from 10 to 50 to assess search depth (Figure 4). Performance generally improves through 40 turns; at 50, EM falls by 4.9 points while F1 rises by 0.5. Continued exploration after finding suitable candidates may drift beyond the relevant event, introduce false positives, or overwrite correct answers. Thus, some queries may gain partial recall while fewer retain exact answer sets. This explanation requires confirmation from paired predictions and traces; the divergence motivates stopping and evidence preservation over larger budgets alone.

Agent Behavior Analysis. We examine tool use across all 122 DISBench queries (Figure 3). ViewPhotos is activated for nearly every query but called more often for inter-event reasoning, suggesting repeated visual comparison rather than broader tool coverage. PhotoEvidence and VisualClues likewise favor inter-event queries, whereas Story tools have similar activation rates across categories and FilterMemoryScope favors intra-event queries. These patterns suggest shared episodic retrieval with task-dependent evidence gathering.

Successful trajectories typically proceed through retrieval, visual inspection, scope verification, and final auditing. Failures include expanding a story before verifying its anchor, transferring anchor-only constraints to targets, and adding candidates without independent verification; premature stopping instead misses relevant photos. These observations motivate separating anchor grounding, scope transfer, and candidate verification.

6 Conclusion

We presented Mneme, a memory-augmented agent for context-dependent image retrieval in personal albums. Its Album Memory Graph provides cross-photo relations, its offline-induced skills guide search decisions, and its dual-route inference reconciles complementary candidates using photo evidence and retained context. Results on DISBench and CamRoll support this combined design across the two tested backbones. The cumulative ablations further show that providing album context alone is insufficient in the tested configuration, motivating joint consideration of memory representations and search procedures. Future work can examine their independent contributions and improve candidate coverage while reducing the cost of graph construction and dual-route inference.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic skill-induction sets for DISBench-style retrieval and CamRoll-style question answering, to improve the clarity and readability of the manuscript. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis. We have reviewed all AI-assisted work. The generated synthetic data were manually inspected and validated by the authors to ensure consistency with the intended task definitions and data-generation requirements. All AI-assisted writing was reviewed and revised by the authors to ensure technical accuracy, clarity, and consistency with the actual methods, experiments, and results reported in the paper. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies personalized image search over long-term photo collections. All data and images used in our experiments are obtained from publicly available research datasets and are used in accordance with their applicable licenses and terms of use. Our auxiliary YFCC-100M albums are used solely to construct synthetic skill-induction sets and are disjoint from the evaluation benchmarks in users, questions, answers, photo identifiers, and trajectories. We do not collect new personal data from users.

Although our Album Memory Graph groups visually consistent face and object instances, these visual clues are used only to associate recurring evidence within an album and are not used to identify real individuals or infer their real-world identities. The intended application of our method is to help users retrieve and reason over their own or otherwise authorized photo collections, rather than for surveillance or cross-user identity tracking.

Accordingly, our data usage and experimental protocol remain within the scope permitted by the corresponding datasets and licenses. We are not aware of additional ethical concerns arising from the methodology or experiments conducted in this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main text describes the architecture of MNEME, memory and skill mechanisms, inference procedure, and experimental setup. The appendix documents Album Memory Graph construction and access tools, synthetic skill-induction set construction, baseline implementations, and the prompts for the agent and free-form answer evaluation. To further support reproducibility, we provide the source code in the Supplementary Materials, enabling reproduction of the proposed method and experimental pipeline.

REFERENCES

Anthropic. Claude opus 4.5 system card. Technical report, Anthropic, November 2025. URL https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel (eds.), *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pp. 758–759. ACM, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

Chenlong Deng, Mengjie Deng, Junjie Wu, Dun Zeng, Teng Wang, Qingsong Xie, Jiadeng Huang, Shengjie Ma, Changwang Zhang, Zhaoxiang Wang, Jun Wang, Yutao Zhu, and Zhicheng Dou. Deepimagesearch: Benchmarking multimodal agents for context-aware image retrieval in visual histories. *CoRR*, abs/2602.10809, 2026. doi: 10.48550/ARXIV.2602.10809. URL https://doi.org/10.48550/arXiv.2602.10809.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Memp: Exploring agent procedural memory. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 17490–17502. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.866. URL https://doi.org/10.18653/v1/2026.findings-acl.866.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

David M. Frohlich, Allan Kuchinsky, Celine Pering, Abbe Don, and Steven Ariss. Requirements for photoware. In Elizabeth F. Churchill, Joseph F. McCarthy, Christine Neuwirth, and Tom Rodden (eds.), *CSCW 2002, Proceeding on the ACM 2002 Conference on Computer Supported Cooperative Work, New Orleans, Louisiana, USA, November 16-20, 2002*, pp. 166–175. ACM, 2002. doi: 10.1145/587078.587102. URL https://doi.org/10.1145/587078.587102.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webwatcher: Breaking new frontier of vision-language deep research agent. *CoRR*, abs/2508.05748, 2025. doi: 10.48550/ARXIV.2508.05748. URL https://doi.org/10.48550/arXiv.2508.05748.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *CoRR*, abs/2601.04720, 2026. doi: 10.48550/ARXIV.2601.04720. URL https://doi.org/10.48550/arXiv.2601.04720.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, Hongkang Yang, Zhi-Qin John Xu,

and Feiyu Xiong. Memos: An operating system for memory-augmented generation (MAG) in large language models. *CoRR*, abs/2505.22101, 2025. doi: 10.48550/ARXIV.2505.22101. URL https://doi.org/10.48550/arXiv.2505.22101.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, Yi Yu, Shuyue Hu, Botian Shi, and Ding Wang. Memverse: Multimodal memory for lifelong learning agents. *CoRR*, abs/2512.03627, 2025. doi: 10.48550/ARXIV.2512.03627. URL https://doi.org/10.48550/arXiv.2512.03627.

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. *CoRR*, abs/2508.09736, 2025. doi: 10.48550/ARXIV.2508.09736. URL https://doi.org/10.48550/arXiv.2508.09736.

Kailin Lyu, Zhiqiang Yuan, Jianwei He, Qiwei Yan, Xuanbo Su, Nanxing Hu, Yang Liu, Ce Hao, Shengqian Qin, Lianyu Hu, Jinchao Zhang, and Jie Zhou. Photocraft: Agentic reasoning with hierarchical self-evolving memory for deep image search. *CoRR*, abs/2606.03099, 2026. doi: 10.48550/ARXIV.2606.03099. URL https://doi.org/10.48550/arXiv.2606.03099.

David Marr, David Willshaw, and Bruce McNaughton. Simple memory: a theory for archicortex. In *From the retina to the neocortex: Selected papers of David Marr*, pp. 59–128. Springer, 1991.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M. Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *CoRR*, abs/2510.12801, 2025. doi: 10.48550/ARXIV.2510.12801. URL https://doi.org/10.48550/arXiv.2510.12801.

Thao Nguyen, Krishna Kumar Singh, Donghyun Kim, Yong Jae Lee, and Yuheng Li. Personal AI agent for camera roll VQA. *CoRR*, abs/2606.05275, 2026a. doi: 10.48550/ARXIV.2606.05275. URL https://doi.org/10.48550/arXiv.2606.05275.

Viet Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal visual memory from explicit and implicit evidence. *CoRR*, abs/2605.28806, 2026b. doi: 10.48550/ARXIV.2605.28806. URL https://doi.org/10.48550/arXiv.2605.28806.

OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. *CoRR*, abs/2509.25140, 2025. doi: 10.48550/ARXIV.2509.25140. URL https://doi.org/10.48550/arXiv.2509.25140.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Kerry Rodden and Kenneth R. Wood. How do people manage their digital photographs? In Gilbert Cockton and Panu Korhonen (eds.), *Proceedings of the 2003 Conference on Human Factors in Computing Systems, CHI 2003, Ft. Lauderdale, Florida, USA, April 5-10, 2003*, pp. 409–416. ACM, 2003. doi: 10.1145/642611.642682. URL https://doi.org/10.1145/642611.642682.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances*

in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

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

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Endel Tulving and Donald Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80:352–373, 09 1973. doi: 10.1037/h0020071.

Yu Wang and Xi Chen. MIRIX: multi-agent memory system for llm-based agents. *CoRR*, abs/2507.07957, 2025. doi: 10.48550/ARXIV.2507.07957. URL https://doi.org/10.48550/arXiv.2507.07957.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wang25bx.html.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 2456–2487. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.114. URL https://doi.org/10.18653/v1/2026.acl-long.114.

Tianyi Xu, Rong Shan, Junjie Wu, Jiachen Zhu, Jiadeng Huang, Teng Wang, Wenteng Chen, Minxin Tu, Quantao Dou, Zhaoxiang Wang, Changwang Zhang, Weinan Zhang, Jun Wang, and Jianghao Lin. Photobench: Beyond visual matching towards personalized intent-driven photo retrieval. In Won Kim, Jae-Gil Lee, Kyuseok Shim, James Caverlee, David F. Gleich, Xiangliang Zhang, Qi He, Faisal Farooq, Rómer Rosales, Sayan Ranu, Haixun Wang, Yan Liu, Carl Yang, Leman Akoglu, Nitesh V. Chawla, and Vagelis Papalexakis (eds.), *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD 2026, Jeju Island, Korea, August 9-13, 2026*, pp. 10092–10103. ACM, 2026. doi: 10.1145/3770855.3817511. URL https://doi.org/10.1145/3770855.3817511.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. *CoRR*, abs/2512.02425, 2025. doi: 10.48550/ARXIV.2512.02425. URL https://doi.org/10.48550/arXiv.2512.02425.

Chenghao Zhang, Guanting Dong, Xinyu Yang, and Zhicheng Dou. Towards mixed-modal retrieval for universal retrieval-augmented generation. In Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (eds.), *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, Melbourne, VIC, Australia, July 20-24, 2026*, pp. 2308–2319. ACM, 2026. doi: 10.1145/3805712.3809716. URL https://doi.org/10.1145/3805712.3809716.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: LLM agents are experiential learners. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 19632–19642. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29936. URL https://doi.org/10.1609/aaai.v38i17.29936.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 19076–19095. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.935. URL https://doi.org/10.18653/v1/2025.acl-long.935.

APPENDIX

A Album Memory Graph Construction Details **15**

B Memory Graph Access Tools Details **16**

C Synthetic Skill-Induction Set Construction **16**

D Baseline Details **17**

E Prompts **18**

E.1 LLM-Based Answer Evaluation Prompt 18

E.2 System Prompt 19

A ALBUM MEMORY GRAPH CONSTRUCTION DETAILS

We construct the Album Memory Graph independently for each user. The graph combines photo-level semantic memories, multi-scale episodic organization, and recurring cross-photo visual associations.

For each photo, we extract available metadata, including capture time and location, and use Qwen3.5 (Team, 2026) to generate a photo-level memory describing salient scenes, people, objects, actions, and visible text. These photo memories are first grouped using temporal, spatial, and semantic cues to form event candidates. Qwen3.5 then assesses the coherence and photo membership of each candidate and summarizes accepted candidates as activity memories. Daily and location memories are constructed from the photo-level evidence, while broader experience memories are synthesized from existing stories and events. This process progressively consolidates individual photo observations into episodic memories at multiple temporal and semantic scales.

In parallel, we extract recurring visual clues across the album. The Buffalo_L model pack from InsightFace detects and embeds faces, while YOLO26 detects object instances. We represent detected object regions with Qwen3-VL-Embedding and cluster the resulting visual representations within each user's album. Qwen3.5 subsequently verifies candidate cross-photo associations for face and object clusters. Each verified association that connects multiple photos forms a visual clue node; singleton face and object detections are excluded from the final visual graph.

The graph contains three functional node families. Photo nodes $\mathcal{V}_P$ represent individual photos and store their stable identifiers, image paths, and capture metadata. Visual clue nodes $\mathcal{V}_C$ represent recurring, cross-photo visual patterns, such as a clustered face or object, and connect to their supporting photos through typed `has_clue` edges. Story nodes $\mathcal{V}_S$ represent episodic context at different scales. They include daily, location, activity, and experience memories, each linked to its supporting photos and, for higher-level stories, to its constituent stories through typed relations. Photo-level captions are retained as textual photo memories associated with the corresponding Photo Nodes and are used to support story construction and caption retrieval.

The resulting graph is stored as per-user Parquet shards. Separate visual and textual FAISS indexes are constructed for photo, clue, and story retrieval. At inference time, the graph loader retrieves the shards and indexes associated with the queried user. Story nodes provide temporal, spatial, and episodic search scopes; clue nodes support recurring visual association; and photo nodes provide direct access to image and metadata evidence. Graph retrieval therefore proposes candidates and reconstructs search scopes, while final photo selection is determined through direct visual and metadata verification.

B MEMORY GRAPH ACCESS TOOLS DETAILS

The agent accesses the Album Memory Graph through seven tools. Their interfaces retain the distinction between contextual proposals, candidate scopes, and directly inspected evidence.

`SearchStoryNodesTool`. retrieves ranked Story Node synopses for event, trip, day, place, or activity queries. It returns no Photo IDs at the synopsis stage, preventing a story label from being mistaken for target evidence.

`SearchPhotosInStoryTool`. expands one exact Story Node internally and searches its bounded Photo Node scope. It supports semantic and lexical include/exclude conditions, pagination, scope counts, and counterfactual candidates when direct target evidence is stronger outside the selected story.

`InspectMemoryPhotoTool`. inspects one Photo Node and returns its metadata, caption memory, connected Story Nodes, connected Clue Nodes, and, when enabled, exact temporal neighbors.

`SearchVisualCluesTool`. searches recurring face and object Clue Nodes. It returns verified associated Photo IDs with pagination information. Clue retrieval proposes visual correspondences but does not replace direct image inspection.

`SearchPhotoEvidenceTool`. performs additive album-wide target recall by fusing a dense Photo Node channel and a caption-memory channel with fixed reciprocal-rank fusion (Cormack et al., 2009). It optionally restricts retrieval to an explicit photo or Story scope.

`FilterMemoryScopeByMetadataTool`. applies deterministic time, address, GPS, and scope operations. For bounded all-matching queries, it also returns a set certificate describing included, excluded, uncertain, and remaining photos.

`InspectCandidateNeighborhoodsTool`. reverse-checks directly observed target candidates by returning their bounded same-day temporal neighborhoods. These neighbors are anchor proposals only and must be visually verified before context is transferred.

C SYNTHETIC SKILL-INDUCTION SET CONSTRUCTION

We construct synthetic skill-induction sets for image-set retrieval and camera-roll question answering using 24 auxiliary YFCC-100M (Thomee et al., 2016) albums. The sets supply queries and reference answers for extracting and refining search skills, rather than serving as held-out validation sets. The auxiliary data are disjoint from the evaluation benchmarks in users, questions, answers, photo identifiers, and trajectories. The construction procedure follows the task formats of DISBench (Deng et al., 2026) and CamRoll (Nguyen et al., 2026a), using automatic generation followed by verification that rejects candidates with failed or inconsistent checks.

Album graph construction and sampling. For each auxiliary user u, we construct an album memory graph $G_u = (V_u, E_u)$, whose nodes represent photos, visual clues, and structured story memories, and whose edges encode caption, clue, story, and photo-membership relations. We first select a bounded clue or story inventory containing between 2 and 64 photos. Starting from this inventory, we perform a seeded balanced random walk across edge families and sample at most 40 edges. Every selected clue or story node is then expanded to its complete photo inventory. As a result, each sampled subgraph defines an explicit reasoning scope $S \subseteq V_u$ for which the relevant photo inventory is known to be complete. Importantly, graph sampling is performed exclusively over the auxiliary albums and does not inspect any target-benchmark questions or answers.

DISBench-style synthesis. For each sampled subgraph, we use `Claude-Opus-4.5` (Anthropic, 2025) to generate a bounded pool of candidate retrieval relations. The generator receives only the sampled graph and is instructed to produce a relation specification rather than a trusted

answer label. Each candidate specifies one distinctive anchor photo, a set of target photos, a context predicate, a context-free visual retrieval predicate, a complete answer scope, and at least one proposed visual hard negative. A candidate must satisfy the following structural constraints: (i) the anchor is unique within the supplied evidence; (ii) the anchor is explicitly excluded from the answer set; (iii) the target predicate is defined by concrete visual properties; (iv) the context-free predicate retrieves both target and hard-negative photos; and (v) the contextual relation separates the target photos from the hard negatives. Generator-proposed answer photo IDs are retained only for diagnostic provenance and are never treated as final labels.

Let $\mathcal{C}$ denote the generated candidate pool. Each candidate $c \in \mathcal{C}$ is independently verified using the deployed Qwen3.5-27B model together with the Qwen3-VL-Embedding retriever. For every declared scope, the verifier exhaustively labels all photos in that scope. Two blind verification passes are conducted with independently shuffled option or label orders and independent request hashes. The verifier additionally checks anchor uniqueness, target–hard-negative separation, target-only retrieval breadth, and whether direct retrieval without the contextual predicate fails to uniquely identify the answer set. If the two scope-level judgments disagree, we invoke a separate adjudication call. Any candidate is rejected if a required verification or adjudication call fails. Query wording may be refined only when an independent semantic-equivalence check confirms that the underlying retrieval relation is unchanged.

We select the final DISBench-style skill-induction set using a constraint solver over candidates that pass all verification gates. The resulting set contains 72 questions from 24 auxiliary albums, with exactly three questions per album and a balanced split of 36 intra-event and 36 inter-event questions. Quality scores are considered only after all distributional and per-user feasibility constraints have been satisfied.

CamRoll-style synthesis. For CamRoll-style questions, we construct an evidence pack for each album from structured story memories, key moments, representative photos, and direct photo-evidence descriptions. Private graph and photo identifiers are replaced with stable local aliases before being exposed to the generator. Using `claude-opus-4.5` (Anthropic, 2025), we separately generate semantic and episodic multiple-choice candidates. A semantic question must describe a stable preference, habit, or recurring pattern supported by at least two independent story scopes. An episodic question must be grounded in a concrete event, location, temporal relation, or visual detail, and must declare one of three reasoning scopes: single-image, multi-image, or whole-roll. Each candidate contains one concise correct answer and two album-aware distractors matched in answer type, grammatical structure, and approximate length.

Candidate options are then verified using `claude-opus-4.5` through two blind full-album audits. For each option, the verifier independently retrieves supporting album evidence with the Qwen3-VL-Embedding index and checks answer support, visual grounding, distractor contradiction, and consistency with the declared reasoning scope. The two audits use different option orders and blinded request contexts. A candidate is retained only if both audits agree and all required evidence checks pass. In particular, semantic questions are rejected when the claimed pattern is supported by only a single event, while episodic questions are rejected when the declared photo evidence does not directly support the answer.

The final CamRoll-style skill-induction set contains 96 questions, with four questions per auxiliary album. Its aggregate composition is fixed at 77 episodic and 19 semantic questions, including 21 detail and 75 non-detail questions. These counts are the integer-scaled counterpart of the distribution observed in the 200-question CamRoll sample. Selection is performed exclusively over verified candidates and does not use any target-benchmark questions, answers, predictions, or scores.

D Baseline Details

Naïve RAG. We implement a simple retrieval-augmented baseline that performs a single-stage retrieval over the user's complete photo album without iterative exploration or persistent memory. Specifically, we use Qwen3-VL-Embedding-8B to retrieve the top-10 images using the original question as the retrieval query, and provide the retrieved images directly to the backbone model for prediction.

PhotoCraft (Lyu et al., 2026). PhotoCraft is a training-free hierarchical memory framework that augments a photo-search agent with working, episodic, and semantic memory. Its semantic memory distills reusable search strategies from successful trajectories and retrieves relevant skills as guidance for subsequent queries. Since no official implementation was available at the time of our experiments, we reimplement PhotoCraft following the method description and released prompts. For comparable supervision, we perform its Skill Distillation on the same synthetic skill-induction set used by our method and freeze the resulting skill library before evaluation. We additionally remove prompt references that explicitly identify the target test set, avoiding evaluation-set-specific information in the baseline instructions.

ImageSeeker (Deng et al., 2026). ImageSeeker is a modular agent for deep image search that combines multimodal retrieval, metadata operations, and direct visual inspection with explicit state and compressed context memory for multi-step navigation over visual histories. We use the official implementation and retain its original tools, memory mechanisms, and inference procedure, replacing only the backbone model with the corresponding model used in our experiments.

CamRoll-Agent (Nguyen et al., 2026a). CamRoll-Agent organizes a personal camera roll into a hierarchical, queryable memory over events, captions, and images, and navigates this memory through a compact set of retrieval, metadata, and visual-inspection tools. We use the official implementation, replacing only the inference backbone and the model used for memory construction with the corresponding backbones in our experimental setup.

E PROMPTS

E.1 LLM-BASED ANSWER EVALUATION PROMPT

We use the following prompt to grade free-form model responses against their gold answers.

Free-Form Answer Evaluation Prompt

```
You are a careful evaluation assistant.
You are grading a model's free-form answer to a visual memory
    question.

Your job is to compare the model response against the gold answer and
    assign a score from 0 to 10.

Scoring rubric:
- 10: Fully correct. Same meaning as the gold answer, even if wording
    differs slightly.
- 7-9: Mostly correct, but less precise or missing a minor detail.
- 4-6: Partially correct. Some overlap, but important information is
    missing or uncertain.
- 1-3: Mostly incorrect, but contains a small relevant overlap.
- 0: Clearly wrong, contradictory, unrelated, or empty.

Evaluation rules:
- Focus on semantic correctness, not wording.
- Accept harmless variation like singular/plural, articles, or minor
    paraphrase.
- For yes/no or preference questions, if the response gets the main
    polarity right but misses intensity or nuance, give medium credit
    rather than a very low score.
- Be strict about specific factual answers: if the response gives a
    different person, place, time, number, yes/no polarity, family
    relationship, or other clearly specified attribute, score 0.

Return ONLY the integer score from 0 to 10.
Do not return JSON.
```

```
Do not return any explanation.

Question: {question}
Gold answer: {gold_answer}
Model response: {model_response}
```

E.2 SYSTEM PROMPT

We use the same system prompt as ImageSeeker, reproduced below for completeness.

MNEME System Prompt

```
You are the Deep Image Search Agent, designed to help users find
    specific photos in their personal photo collections.
Your role is to understand user queries, reason over images and
    metadata, and orchestrate tool calls to retrieve the correct
    photos.

CORE OBJECTIVES
- Your primary goal is to locate the **exact** photos a user has in
    mind.
- You must perform multi-step reasoning: break down complex queries,
    call tools, and locate the target photos.
- Use both visual understanding and metadata to narrow down to the
    most relevant photos.

YOUR TASK

1. Understand the user's query
  - Carefully read the full query, perform explicit query analysis
    and structure it into the following components:
    A. EPISODE (latent context)
      - Infer the complete underlying EPISODE implied by the query.
      - An EPISODE is a single event or group of events involving
    time-spatial movements, coexistence of objects, or logical steps.
    B. EPISODE Breakdown (logical path)
      - Explicitly break down the EPISODE into a detailed
    step-by-step description. Each step should be a photo, or a group
    of photos, or an event.
      - Describe each photo by visual content, metadata constraints
    and other constraints.
      - Explicitly define logical relations among steps, photos, and
    events. Common relations include (but are not limited to): same
    time (or window), same place (or nearby area), same event, same
    people, same entity.
    C. TARGET photos (to be returned)
      - Identify what the TARGET photos to be finally returned must
    depict. Describe TARGET photos along two dimensions: visual
    content and metadata constraints.
      - Treat TARGET descriptions as strict requirements.
      - Identify between anchor photos and TARGET photos. Do NOT
    require anchor-specific objects or scenes to appear in TARGET
    photos.
      - Do NOT require EPISODE evidence to appear in TARGET photos
    unless explicitly stated.

2. Plan a search strategy
  - Decompose the query into logical sub-problems.
```

```
    - Use context references as anchors when helpful, for example to
    infer a relevant date, time window, address, people, or scene
    context.
    - After extracting such constraints, carefully choose the
    appropriate search space (the whole album or a saved subset)
    based on the user's intent and the structure of the query.
  - Design a step-by-step plan that might involve:
    - Searching the whole album with a text or multimodal query.
    - Filtering by time and/or address.
    - Reusing previously saved subset of photos.
    - Inspecting small sets of candidate images to verify
    fine-grained details.
  - If the candidates you see do not contain plausible matches, stop
    and reassess rather than pushing forward with a failing plan.
    Refine the plan before continuing.

3. Use tools iteratively and efficiently
  - Utilize important visual clues in images across steps.
  - Use tools to:
    - Retrieve candidates with high recall.
    - Validate/derive time and address from metadata.
    - Perform visual inspection to confirm target attributes.
  - Events often span multiple days/nearby locations; treat
    dates/addresses as approximate unless the query specifies
    exactness.
  - Be careful when choosing search scope and filter scope. Apply
    scope constraints only when necessary.
  - Perform explicit reasoning before calling tools.

4. Verify candidates before answering
  - For each candidate considered for the final set, verify:
    - It satisfies TARGET visual requirements.
    - It satisfies any TARGET metadata constraints.
    - Anchor constraints are applied only when they define the
    event/context; do not require anchor visuals inside targets
    unless stated.
    - When a query mentions an event in a temporal phrase (e.g. "on
    the day we...", "during the trip to...", or "after we
    visited..."), treat this primarily as a constraint on the time or
    address, not as a requirement that the event must appear in the
    final photos unless explicitly stated.
  - Return all photos that meet the intent.
  - If nothing perfectly matches, return the closest matches.
  - In all cases, you MUST output a clear, explicit list of photo IDs
    as your final answer in the format: "The final answer is:
    ["photo_id1", "photo_id2", ...]."

RULES

- Never fabricate photos, metadata, or any other information. It is
    better to call extra tools and verify than to guess.
- Be careful and methodical in planning, and strategic in combining
    visual clues, metadata, and intermediate subsets.
- When dealing with ambiguous references ("Grandma's house", "our
    road trip"), infer their meaning from context and prior results
    and reuse these inferred identities consistently.
- Do not ask the user questions request confirmation. If information
    is missing or ambiguous, make a best-effort inference based on
    available evidence and proceed.
- Do not offer options such as "Would you like me to refine the
    search further...". Instead, autonomously refine the search if
    needed, then return your best final result.
```

```
- In your final response, you MUST include a single explicit
    statement in the exact format: "The final answer is:
    ["photo_id1", "photo_id2", ...]."
```