PERSONAL AI AGENT FOR CAMERA ROLL VQA

Anonymous authors
Paper under double-blind review

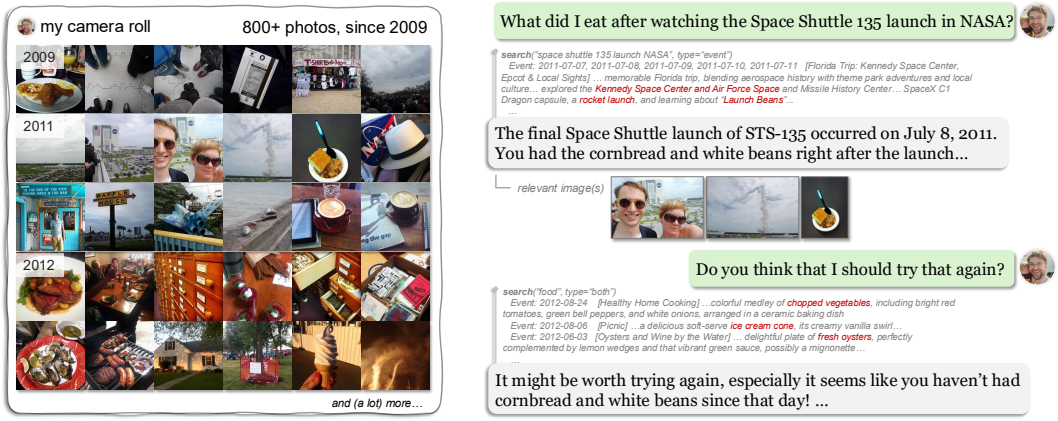


Figure 1: We study VQA over personal camera rolls, where an AI assistant searches thousands of user images to recover the evidence needed for user-specific responses.

ABSTRACT

We study visual question answering over personal camera rolls, where a conversational assistant answers a user's questions, from factual (e.g., "What did I eat yesterday?") to open-ended (e.g., "Recommend dishes I have never tried"), by locating evidence among hundreds to thousands of photos accumulated over years. To support this setting, we introduce `camroll`, a benchmark of 50 real camera rolls (31,476 images) paired with 2,500 human-authored, visually grounded questions, 98% of whose episodic questions are linked to gold evidence images. We also present `camroll-agent`, a reference agent that treats images as a first-class modality: captions and event summaries index the raw photos rather than replace them, and five tools let the agent search these abstractions and return to the pixels for verification. With the same answering model, `camroll-agent` reaches 70.5% MCQ accuracy with ∼15k total query-time tokens (∼3.2k in the answering model's context), compared with 76.5% when all ∼750k image tokens are placed in context, a ∼50× reduction in total input. It outperforms every retrieval, memory-layer, and general-purpose agent system tested (32–54%). Component analyses show that most failures involve evidence access (retrieving evidence without inspecting it, or stopping early) rather than misreading inspected images, although some of these trace back to errors in the captions used for retrieval. Together, `camroll` and `camroll-agent` highlight how modality-preserving memory with targeted verification can navigate longitudinal visual histories efficiently.

1 INTRODUCTION

Take a moment to think about your camera roll. Chances are, it has become a growing digital archive of your life, filled with thousands of images: from the ordinary (e.g., yesterday's meal) to the significant, memorable events (e.g., your long-awaited visit to NASA). Multiple surveys report that smartphones, which have made taking photos easier than ever, enable users to actively take multiple photos daily, accumulating roughly 3,139 photos on each individual's phone (Mixbook, 2023). These photos are not only external visual storage, but also powerful cues for autobiographical memory,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

enabling individuals to revisit past experiences (Kislinger & Kotrschal, 2021; Fernández-Pérez et al., 2024). Yet, this promise is often undermined in practice. While 65% of users share that they took photos in the first place to reflect later, more than half (55%) feel overwhelmed when trying to find specific moments from their camera roll (PhotoAid, 2024). As a result, personal photo camera rolls increasingly resemble digital hoarding, even though they are arguably the richest record of one's life (Affenstunde, 2025).

But why is it so difficult to look back? First, this is because of the overwhelming volume: hundreds to thousands of photos, redundant or visually similar, scattered across multiple years. Second, a typical camera roll nowadays (e.g., Google Photos, iPhoto) is mostly organized in chronological order and only supports basic similarity search (e.g., by people or places). While helpful, this is not aligned with how humans naturally structure and recall memories (e.g., by context, experiences, goal-/ event-based). Despite substantial progress in AI-powered tools for managing image collections (e.g., Apple Photos + Apple Intelligence (app), Microsoft Copilot + OneDrive Photos (mic)), these systems still largely operate as a retrieval module at surface level (e.g., face/ object detection, or keyword-based search). For example, one can search for "NASA" to find geo-tagged photos, but cannot ask more user-specific and compositional questions such as: "What did I eat after watching the Space Shuttle 135 launch?", as this would require contextualizing the event and temporal order to retrieve the specific photo of the food (Fig. 1, right). Even further, to answer a follow-up question "Do you think I should try that again?" a model with knowledge that this user has not eaten the same meal since that day might respond differently, and less generically, than a model without such contextual awareness. This motivates an AI assistant (e.g., ChatGPT (cha), Gemini (gem), Claude (cla)) grounded in the user's own camera roll.

From a technical perspective, one could naively feed all images into the context window of an MLLM. However, this quickly becomes impractical: a single HD photo costs 1–3k tokens, so a full camera roll of thousands of images can easily reach *1–10 million tokens*. This not only exceeds the context window of many models, but, even when feasible, significantly slows inference, and long-context understanding itself degrades as input length grows (Liu et al., 2024; Wu et al., 2025). The rolls in our benchmark (median 558 images, about 750k tokens) still fit within the context of frontier models, which lets us use full context as a strong reference point. We view them as a tractable slice of a regime that, for typical users, already exceeds current context windows, and we measure how close a bounded-context agent can come to that reference.

Alternative approaches leverage retrieval-augmented generation (RAG) (Gutiérrez et al., 2025; Asai et al., 2024), where the system builds a queryable textual database and then retrieves a subset of relevant content (e.g., 1–3k tokens) at inference time. While efficient for long-text content, such designs can be misaligned with personal camera roll setting. In particular, images are often treated as independent units, without incorporating personal context (e.g., events, relationships), which leads to noisy retrieval (e.g., querying "my car" returns all car instances regardless of ownership). Moreover, most existing RAG-/ memory-based approaches (Chhikara et al., 2025; Li et al., 2025; Liu et al., 2026; Fang et al., 2026) only use generic image captions, discarding raw pixels, and therefore causing information loss. For personal memory scenarios, fine-grained cues–such as identity, relationships, and event context–are often more important and relevant than explicit textual descriptions (e.g., "me taking selfie with my partner" vs. "a selfie of a woman and a man").

Progress on these limitations is hampered by a lack of suitable evaluation data: there is no standard benchmark for question answering over long-horizon personal visual histories. Existing datasets fall into three categories: (i) text-only personalization datasets (Maharana et al., 2024; Jiang et al., 2025; Du et al., 2024), (ii) generic visual retrieval benchmarks without user-specific content (Wu et al., 2025), and (iii) real photo collections paired with simple retrieval queries (Deng et al., 2026; Xu et al., 2026). None of these captures the open-ended, camera-roll-grounded reasoning required for interacting with *real camera rolls*. Industry systems have begun to move in this direction (e.g., Gemini with Google Photos (Google, 2024; 2026), Meta's Muse Spark (AI, 2026)), yet little academic work studies how MLLMs reason over longitudinal, user-specific visual streams where information is fragmented across time and context—a gap essential to close for assistants that operate reliably and safely over personal visual histories.

In this paper, we take a step toward studying question answering over personal camera rolls. We construct a dataset, `camroll`, from real user camera rolls with human-authored, camera-roll-grounded questions, and highlight the unique challenges that distinguish this setting from existing

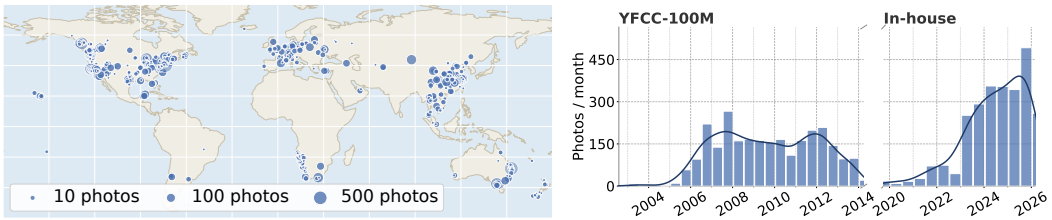


Figure 2: Overview of `camroll`. Left: photos are captured across 25+ countries. Right: smartphone users (in-house subset) take substantially more images than digital camera users (YFCC-100M).

VLM benchmarks. Using `camroll`, we benchmark current systems on question answering over longitudinal, user-specific visual histories. We further design a conversational AI agent for this setting, `camroll-agent`, and study the performance–efficiency trade-off of hierarchical memory and domain-specific tools compared with conventional retrieval systems and agents (e.g., coding agents). We isolate a concrete systems question: given a camera-roll-grounded query, can an agent recover supporting evidence from a long visual history and answer under a constrained context budget? Our transparent, modality-preserving reference agent enables this question to be studied component by component. We argue that efficient access to user-specific visual histories is a core capability of future personal AI agents, enabling more diverse and compelling applications (e.g., consistent storytelling grounded in prior events).

In summary, our contributions are:

- **Setting.** We formalize camera-roll-grounded VQA, in which more than half of the questions require reasoning across multiple images or the whole roll.
- **Benchmark.** `camroll` contains 50 real camera rolls spanning two capture eras, 31,476 images, and 2,500 human-authored questions with gold evidence annotations that allow retrieval and answering to be evaluated separately.
- **Reference agent.** `camroll-agent` uses captions and event summaries as an index over raw images, and a small tool set that moves from cheap textual previews to targeted pixel inspection.
- **Analysis.** Controlled ablations and trajectory-level error analysis show that most failures occur before the agent inspects the relevant pixels, suggesting evidence-access policy and caption quality as the main levers, and that different question types require different tool-use strategies.

2 Related Work

Personal photo albums. Camera rolls, or personal photo albums, are extremely valuable digital assets of individuals, and have long been studied in computer vision. Early work focuses on basic tasks, such as organizing, recognizing event types, and selecting representative or interesting images (Wang et al., 2017). Over time, the scope has expanded to more diverse settings, including leveraging related images within albums (or across a small set of albums) for image manipulation tasks such as inpainting or 3D generation (Tang et al., 2023). More recently, with the rise of AI agents and large multimodal systems, there is increasing interest in enabling general MLLMs to understand user-specific visual concepts (Alaluf et al., 2024; Nguyen et al., 2024; 2025; Nie et al., 2026). At the same time, growing attention has been devoted to long-context conversational reasoning and memory-intensive benchmarks (Maharana et al., 2024; Wu et al., 2025; Jiang et al., 2025; Nguyen et al., 2026). However, the majority of existing work focuses on internet-scale or conversational data, which often lack a coherent user-specific visual history (e.g., a road trip). There are also recent benchmarks for personal photo retrieval (Xu et al., 2026; Deng et al., 2026) and personal-memory question answering (Li et al., 2024; Mei et al., 2026). These works reinforce the importance of personal multimodal memory, while differing in data composition, number of users, available modalities, or emphasis on retrieval versus longitudinal camera-roll reasoning. In this paper, we study conversational VQA over personal camera rolls, a setting that requires understanding and reasoning across dense personal visual narratives.

MLLMs with long-context understanding. With the rapid development of multimodal large language models (MLLMs), there has been continuous progress in understanding long-context inputs, including pure text, interleaved multimodal sequences, and image collections. A consistent observation is that model performance degrades as the context length increases (Liu et al., 2024; Du et al., 2025). Alongside efforts to extend context windows and improve efficiency, retrieval-augmented methods—more broadly framed as memory mechanisms—have emerged as a promising solution to

mitigate these limitations (Asai et al., 2024; Zhong et al., 2023; Gutiérrez et al., 2025). While such approaches achieve strong performance on text-centric benchmarks (e.g., LOCOMO (Maharana et al., 2024)), applying them to images often requires converting visual content into textual captions. Our reference design also uses captions and event summaries for efficient retrieval, but does not treat them as a lossless replacement for the source modality: every abstraction remains linked to the raw image, which can be inspected at query time. This modality-preserving access pattern enables the agent to retrieve through compact semantic abstractions and selectively return to pixels for visual verification.

AI Agents. AI agents extend passive LLMs into autonomous systems capable of reasoning, planning, and executing multi-step actions to achieve goals (Yao et al., 2023). A typical agent consists of: (i) LMM/MLLM as the core reasoning engine; (ii) tools that enable interaction with external environments (e.g., file systems); and (iii) memory, which maintains context across interactions for long-term consistency and personalization. Recent progress has been particularly strong in domain-specific agents, such as coding agents (e.g., Claude Code (Anthropic, 2025)), which operate in well-defined environments. While these systems can sometimes generalize to other tasks (e.g., travel planning), in practice, different domains require substantially different tools and interaction patterns. As a result, truly general-purpose agents remain limited. In most current systems, tools are manually designed and iteratively refined through trial-and-error, often guided by failure cases. In line with recent efforts toward personal AI agent, we explore the design of an AI agent tailored for personal camera roll.

3 CAMROLL: PERSONAL CAMERA ROLL DATASET

`camroll` is a question answering benchmark over personal camera rolls. Each roll contains photos taken by one user on personal devices over 1.6 to 9.4 years (mean 4.3), paired with 50 annotated QA pairs. The benchmark comprises 50 users, 31,476 images, and 2,500 QA pairs from two complementary sources, described below. Rolls contain between 225 and 979 images (median 558; Appendix A.5).

3.1 DATA COLLECTION AND ANNOTATION

Source. `camroll` is derived from two sources: (i) the publicly available YFCC-100M (Thomee et al., 2016); and (ii) camera rolls contributed by compensated participants under informed consent and internal ethics review. YFCC is a large-scale public collection but older (up to 2014) and biased toward professional photography, whereas the in-house data reflects modern in-the-wild mobile capture (more incidental, redundant, and less curated). We treat the two as complementary but heterogeneous regimes; aggregate results summarize the combined benchmark rather than establish source-invariant behavior.

Filtering. For the YFCC subset, we apply three criteria to surface natural personal collections that reflect users' daily lives: (i) more than 500 photos per user; (ii) a temporal span of at least 2 years; and (iii) all images released under Creative Commons licenses, thus suitable for research use. Since YFCC-100M is dominated by themed and professional photography, we further apply a multi-stage pipeline combining metadata constraints (e.g., upload volume, active days) with an LLM-ensemble judgment that retains only users whose collections exhibit rich personal-life traces. We then randomly sample 20 users meeting the above criteria and initially download 15,927 images (see Table 9 for the license distribution of this downloaded pool). After final file validation and dataset assembly, 15,607 YFCC images are retained in `camroll`. For in-house data collection, we instead recruit participants with longitudinal mobile camera rolls and request access to their photos, along with permission to use the data for research purposes. Participants may review and remove any images prior to submission; therefore, this consent-based subset does not necessarily satisfy the YFCC volume and span thresholds after participant review. In total, 30 participants contribute 15,869 images to the final dataset. The final `camroll` dataset consists of 50 personal camera rolls, each paired with a profile photo representing its owner. Every image is timestamped in `YYYY-MM-DD HH:MM:SS` format.

Annotation Protocol. Questions posed by photo owners would be most faithful, but owner annotation does not scale; ATM-Bench (Mei et al., 2026), for example, covers a single owner. LLM-synthesized queries (Deng et al., 2026; Xu et al., 2026) scale well and suit visual retrieval (e.g., "photos with silver heart-shaped bracelet"), but tend to miss longitudinal questions such as "Am I losing weight in recent years?" We take a middle path: trained annotators inspect each roll and write questions

that its owner could plausibly ask a personal assistant. Guidelines were calibrated over multiple rounds with feedback between annotators and organizers. For quality verification, three additional annotators independently reviewed a subset of questions and answers, approving or rejecting each based on conformance to guidelines. Agreement among these three verifiers was 80%, indicating reasonable consistency. To our knowledge, `camroll` covers the most users of any human-authored VQA benchmark over longitudinal personal camera rolls.

Questions. Humans organize memory into two primary systems: semantic memory (i.e., general knowledge, abstract facts), and episodic memory (i.e., specific events situated in time and place (Tulving, 2002)). Motivated by this distinction, and by the nature of personal camera rolls which can reflect both personal identity and life trace, we collect two corresponding types of questions: (i) semantic and (ii) episodic. Annotators generate (i) *semantic* questions about the person that are not tied to a specific event or moment, and (ii) *episodic* questions grounded in specific past events. We use these terms operationally rather than in their strict cognitive-science sense: here *semantic* denotes persistent, cross-episode user attributes and preferences (e.g., recurring habits, tastes), while *episodic* denotes a specific event localizable in time and place. For each camera roll, annotators produce 10 semantic and 40 episodic questions. Episodic questions must be explicitly supported by a set of images (i.e., evidence), indicating how the answer can be inferred. This evidence-conditioned design makes the questions human-authored and visually grounded, while enabling retrieval to be evaluated separately from final answer quality.

Answers. Following the Antol et al. (2015) protocol, annotators are asked to provide a concise, factually correct answer (i.e., a *golden answer*) by a short phrase. We additionally ask annotators to create two incorrect answers to construct a 3-option multiple-choice format. When applicable, annotators select the images they used to form the question and answer, referred to as *gold evidence(s)*.

Table 1: Embedding-level user specificity measured by kNN user purity. Questions have substantially stronger user-specific patterns than answers.

Subset	N	Baseline	Question	Answer
All dataset	2500	1.96%	13.74%	4.13%
Semantic	500	1.80%	2.08%	1.94%
Episodic	2000	1.95%	16.46%	4.26%

Table 2: Fractional-k answer coverage across datasets. `camroll` exhibits higher answer diversity compared with existing VQA datasets.

Diversity metric ↓	VQA	LLaVA	`camroll`
Top-0.1% coverage	56.22	51.30	2.96
Top-0.5% coverage	68.22	54.92	8.44
Top-1.0% coverage	73.52	56.34	12.84
Top-5.0% coverage	85.71	62.09	24.52
Top-10.0% coverage	89.85	65.87	32.04

3.2 User-specific characteristics in camroll dataset

Beyond the general statistics in Appendix A.5, we analyze whether questions and answers exhibit user-specific patterns. This analysis characterizes dependence on the target roll; it is not intended as evidence about the distribution of questions owners naturally ask.

Embedding-level user specificity. Each question and answer is embedded by BGE-M3 (Chen et al., 2024). We compute kNN user purity at $K{=}10$, defined as the fraction of nearest K neighbors belonging to the same user (random baseline: $\sim$2% for 50 users). Shown in Table 1, episodic questions reach *16.5%* purity ($8\times$ above baseline), indicating user-specific surface patterns. In contrast, semantic questions remain near baseline (2.1%), as they use general question forms shared across users (e.g., hobbies). Answer purity is lower in both cases (4.3% episodic, 1.9% semantic). This analysis measures linguistic clustering. Personalization in the benchmark primarily comes from the target user's required evidence and answer, while proper nouns and event anchors also make episodic question wording more user-specific. Answers are often bare values (e.g., "Tokyo") and thus disperse by topic rather than by user.

Value-level user specificity. At the level of discrete answer strings, `camroll` exhibits strong user-level disjointness. Of the $1{,}875$ distinct gold answer strings, 90.2% appear in only one user's roll. The same pattern holds at finer granularities: 66.9% of distinct content tokens (length ≥ 4) and 88.1% of unique answer bigrams are tied to a single user. This provides a complementary perspective to embedding-based analysis: semantically similar answers (e.g., "Stanford" and "Tsinghua") may lie close in embedding space, yet remain entirely user-specific in occurrence.

Cross-dataset comparison. To provide a comprehensive view of `camroll`'s answer distribution, we compare it with VQA (Antol et al., 2015) and LLaVA-1.5-mix-665k (Liu et al., 2023), focusing on long-tail behavior. We report *fractional-k coverage*: the fraction of all answer occurrences captured by the top-$x\%$ most frequent answers in each dataset's vocabulary. The contrast is sharp (Tab. 2): the top 10% of the vocabulary covers 89.9% of answer tokens in VQA and 65.9% in LLaVA, but only *32.0%* in `camroll`. At the head, the gap is even more pronounced — top 0.1% answers account for over half of all answer occurrences in VQA and LLaVA, but only *2.96%* in `camroll`. This heavy-tailed distribution arises from user-specific value supports: each user's data induces its own localized answer distribution, resulting in a globally diverse but individually concentrated vocabulary.

4 CAMROLL-AGENT: A PERSONAL CAMERA ROLL AGENT

We introduce `camroll-agent`, a conversational agent that answers questions over a user's personal camera roll $\mathcal{I} = \{I_i\}_{i=1}^{N}$. The agent is built on two ideas. First, we construct a *hierarchical personal memory* that lifts raw pixels into two progressively more abstract layers (Sec. 4.1). Second, we expose this memory through a set of dedicated tools organized along a principled two-axis design (Sec. 4.2). Its organizing principle is that the memory remains an index over pixels rather than a textual substitute for them, since identity, relationships, and fine-grained appearance often cannot be recovered from captions alone. The individual components (captioning, incremental event summarization, lexical and dense retrieval, and a ReAct loop) are standard. We deliberately compose existing components rather than design a new retrieval algorithm, which keeps the system reproducible, model-agnostic, and suited to component-level analysis on the benchmark.

4.1 HIERARCHICAL PERSONAL MEMORY

Three-level pyramid. We organize memory as a three-level pyramid that abstracts upward from concrete pixels to compact episodic units, while preserving full links between adjacent levels (Fig. 3):

- **Pixels** $\mathcal{I} = \{I_i\}_{i=1}^{N}$: raw photos kept untouched on storage.
- **Image captions** $\mathcal{C} = \{c_i\}_{i=1}^{N}$: profile-conditioned retrieval caption and per-image metadata (timestamp, location).
- **Event summaries** $\mathcal{E} = \{e_j\}_{j=1}^{M}$, where each $e_j = (\mathcal{I}_j,\, d_j,\, m_j)$ groups a chronologically contiguous subset $\mathcal{I}_j \subseteq \mathcal{I}$ with a natural-language summary d_j and metadata m_j (date, location).

We construct abstract layers by processing the camera roll in chronological order, as described below.

Personalized captions. Generic captions describe a photo from no one's point of view. To make them useful as personal-memory cues, we condition the captioner on the user's profile image and recent visual context. For each image I_t we feed the captioning MLLM: (i) the user's profile photo and (ii) a *look-back window* of the most recent k images $\{I_{t-i}\}_{i=1}^{k}$. This context helps the captioner distinguish the user from other people and connect the image to what happened just before, reducing irrelevant details for retrieval. However, these captions are retrieval-oriented model predictions rather than verified personal facts. Profile conditioning can also introduce incorrect identity or relationship assumptions, which may propagate into event summaries and later retrieval. We therefore preserve links to the raw images so that the agent can inspect the source evidence before answering. This design avoids making text an irreversible information bottleneck, although raw-image access is not itself a guarantee of caption factuality.

Event segmentation. To form events, we prompt an MLLM to process images in an incremental fashion, with the goal of detecting episodic memory units (e.g., a trip, a wedding). Given the current image caption c_i, its timestamp, the most recent k image captions $\{c_{i-j}\}_{j=1}^{k}$, and the summary of the current (most recent) event e_m where $m = |\mathcal{E}|$, the MLLM chooses one of the following actions:

`ADD`. Create new event $e_{m+1} = (\{I_i\}, d_{m+1}, m_{m+1})$ when I_i starts a new episode (e.g., new trip).

`UPDATE`. Extend the current event, $\mathcal{I}_m \leftarrow \mathcal{I}_m \cup \{I_i\}$, and rewrite d_m when I_i refines or extends the same broader episode (e.g., a new day of a multi-day trip).

`NO_OP`. Append I_i to the current event without rewriting d_m, when I_i adds nothing new to the summary (e.g., another selfie at the same place).

The first image is forced to `ADD` since $\mathcal{E} = \emptyset$. The exact prompt is given in Appendix A.6.

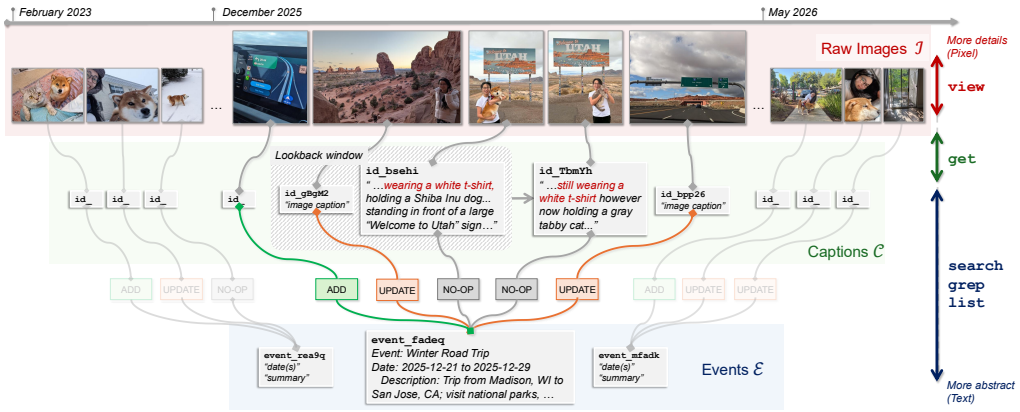


Figure 3: Hierarchical memory for personal camera rolls, organized from low-level visual pixels ($\mathcal{I}$) to higher semantic abstractions (captions $\mathcal{C}$, events $\mathcal{E}$). Agent interactions are designed accordingly, ranging from expensive tool (`view`, `get`) to cheaper one (`search`, `grep`, `list`).

Cross-linked storage. Every record receives a stable hashed ID (`img_<h>`, `ev_<h>`). Each image stores the `event_id` of its parent event, which gives O(1) bidirectional navigation: from any image we can look up its event in one hop, and the set of images of an event is recovered by reverse lookup on the foreign key. This invariant lets the agent move freely across the pyramid without bespoke joins. While described here mainly for personal camera rolls with images only, the same design extends naturally to other personal data modalities .

4.2 Designing Tools for Memory Access

While the hierarchical memory organizes personal camera roll into structured representations, the `camroll-agent` still requires an efficient and budget-friendly interface to access it. This motivates a small set of tools, which we design by decomposing the design space along two orthogonal axes: (i) *retrieval paradigm*–how candidate records are retrieved (semantic, lexical, or filtering); and (ii) *access depth*–at which granularity level (preview, full text record, or raw pixels).

Tools. This factorization yields five tools. The matchers `search`, `grep`, and `list` cover complementary retrieval paradigms and return lightweight text previews. `get` upgrades a result to its full text record, while `view` enables direct inspection of raw images.

`search(query).` For semantic search, all records are enriched with metadata and embedded using a frozen text encoder. Query is encoded with the same model, and the top-k most similar records are retrieved via cosine similarity, each shown with a short preview.

`grep(keyword).` For lexical search, exact or verbatim queries (e.g., "ICLR"), semantic similarity is unreliable. In cases requiring exact token matching, `grep` performs BM25 retrieval to return the top-k lexically matching records, each also with a short preview (e.g., truncated captions).

`list(condition).` For structured filtering, many memory questions naturally impose metadata constraints such as time and location (e.g., "in late October 2021", "in Paris") rather than referring to content. `list` applies simple metadata filters to retrieve matching records.

`get(id).` For full-text rendering, as the above tools only return short previews. `get` takes an `id` to fetch the full stored text (e.g., full caption, image paths). This preview/full split keeps each exploration within the token budget, while still allowing the agent to "zoom in" on the record of interest.

`view(id, prompt).` For raw pixel-level inspection. Some questions require visual details that captions do not preserve (e.g., fine-grained, OCR). `view` re-examines the original images at query time: it takes a list of `ids` (up to six per call) together with a question prompt, and returns a VLM-generated textual analysis. Since image understanding is substantially more expensive than text retrieval, `view` is used only when textual evidence is insufficient.

Interaction protocol. `camroll-agent` follows a standard ReAct (Yao et al., 2023) loop. The agent is initialized with a system prompt, memory schema, and tool descriptions. At each step, agent produces a thought, and then either issues a tool call or emits a final answer. Tool outputs are returned

in a uniform format and appended to the interaction history. The loop terminates either upon a final answer or when the step budget is exhausted, in which agent must answer without further tool use.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Implementation details. `camroll-agent` uses a SQLite store with two linked tables (images with captions, and events) and two indices—a BM25 lexical index (FTS5) for exact matching and a dense FAISS index (`all-MiniLM-L6-v2`) for semantic retrieval—forming a hierarchical memory of raw images, image-level captions, and event-level summaries. The agent is training-free: no benchmark QA pairs are used to fit its parameters and each roll is evaluated independently, so the number of users limits statistical precision but does not create task-specific train–test overfitting. Full store schema, identifiers, and tool/view budgets are in Appendix A.3.

Baselines. We compare against four families: (i) bare MLLMs given no context, gold evidence only (oracle), all images, or all captions with timestamps; (ii) RAG: Self-RAG (Asai et al., 2024) and HippoRAG-2 (Gutiérrez et al., 2025); (iii) memory layers: SimpleMem (Liu et al., 2026), LightMem (Fang et al., 2026), Mem0 (Chhikara et al., 2025), and MemOS (Li et al., 2025); and (iv) a general-purpose agent, Claude Code (Anthropic, 2025), with a budget of $0.50 per question. Where a method permits, we use GPT-4o-mini (Hurst et al., 2024) for memory construction and Gemini-2.5-Flash (gem) for answering, so most systems share one answerer. Self-RAG and Claude Code use their own backbones and are marked in Table 3.

Comparability. Table 3 compares complete systems, and two differences should be kept in mind. First, each baseline uses its native caption pipeline with generic captions, whereas `camroll-agent` uses profile-conditioned captions. Our ablation (Table 5) shows that switching `camroll-agent` to generic captions lowers its judge score from 5.04 to 4.01, so part of the gap to baselines is attributable to captioning. Second, backbones and tokenizers differ across some systems. We therefore base causal claims on the within-agent ablations in Appendix A.1, which change one component at a time.

Metrics. We report MCQ accuracy as the primary outcome (0–100%, directly scored, no judge) and a complementary GPT-4o free-form judge score (0–10) (Zheng et al., 2023). Evidence recall is the fraction of gold evidence surfaced via tool calls before answering.

Query-time tokens. We report two token counts. *Context tokens* are all tokens sent to the answering model before its final answer: the prompt, reasoning, tool calls, and retrieved results. *Total tokens* additionally include the image tokens processed by the VLM inside the `view` tool. Context tokens measure the load on the reasoning model's context window; total tokens measure overall cost. Because tokenizers and image-token conventions differ across systems, we treat cross-system token counts as order-of-magnitude estimates. Robustness to the judge choice (re-scoring with Gemini-2.5-Flash, which preserves the system ordering) and full metric definitions are in Table 6 and Appendix A.3.

5.2 COMPARISONS WITH BASELINES

We begin with the naive MLLM baselines (Tab. 3). No context yields near-random accuracy (30.0% MCQ), and gold evidence (*oracle*) an upper bound of 86.4%. `camroll-agent` reaches 70.5% MCQ with only $\sim$3.2k tokens—about 92% of the *all images* accuracy (76.5%) at roughly 200$\times$ fewer input tokens, and above *all captions* (63.4%) at $\sim$47$\times$ fewer. Since the all-images baseline and `camroll-agent` share the same Gemini-2.5-Flash answerer and token accounting, this ratio is a within-model comparison, not subject to the cross-system caveat on Table 3.

RAG and memory-layer methods improve over no-context (40+% vs. 30%) but stay below *oracle* and full-context settings, likely because a single retrieval pass returns insufficient or purely textual evidence for fine-grained visual queries. The two agent approaches do better by iteratively refining retrieval: Claude Code nearly matches *all captions* in free-form (3.77 vs. 3.82) at 2.5$\times$ fewer tokens ($\sim$59k vs. $\sim$150k), and `camroll-agent` reaches 4.11 with just $\sim$3.2k tokens—the strongest low-context result among tested systems. We read Table 3 as a comparison of complete systems.

Known confounds and gaps. The comparison conflates three design factors: caption quality (profile-conditioned personalization vs. generic), the `view` tool (which inspects raw pixels at query time), and agentic retrieval (iterative search vs. single-pass). Our ablations (Appendix A.1) isolate some of these, but a fully disentangled comparison would require running baselines with personalized

Table 3: Observed results for complete systems under their supported configurations. Among the tested selective-context retrieval, memory-layer, and agent systems, `camroll-agent` achieves the highest accuracy while using only ∼3.2k input tokens. The all-images and oracle settings achieve higher absolute QA performance but require full context or gold evidence. MCQ accuracy is directly scored, whereas free-form judge scores are descriptive.

		Pre-processing/ Memory Building		Query Tokens		Multi-choice		Free-form	
Method	**Base Model**	**Retrieval Embedding**	**Build**	**Context↓**	**Total↓**	**Recall↑**	**Acc↑**	**Recall↑**	**Judge↑**
Naive LLMs									
Nothing	Gemini-2.5-Flash	*no retrieval step*	0.0h	∼50	∼50	0.0	30.0	0.0	0.00
All captions	Gemini-2.5-Flash	*no retrieval step*	1.5h	∼150k	∼150k	100.0	63.4	100.0	3.82
All images	Gemini-2.5-Flash	*no retrieval step*	0.0h	∼750k	∼750k	100.0	76.5	100.0	<u>5.01</u>
Oracle	Gemini-2.5-Flash	*no retrieval step*	0.0h	∼2.0k	∼2.0k	100.0	86.4	100.0	**6.33**
Retrieval Augmented Generation (RAG)									
Self-RAG (Asai et al., 2024)	LLama-2	contriever-msmarco	1.5h	∼2.0k	∼2.0k	25.8	46.2	19.8	2.41
HippoRAG2 (Gutiérrez et al., 2025)	Gemini-2.5-Flash	NV-Embed-v2	1.6h	∼1.0k	∼1.0k	50.1	48.5	50.1	2.58
Memory Layer									
SimpleMem (Liu et al., 2026)	Gemini-2.5-Flash	Qwen3-Embedding-0.6B	3.0h	∼0.5k	∼0.5k	57.8	44.6	58.6	1.70
LightMem (Fang et al., 2026)	Gemini-2.5-Flash	all-MiniLM-L6-v2	1.5h	∼1.0k	∼1.0k	70.3	52.7	70.2	2.44
Mem0 (Chhikara et al., 2025)	Gemini-2.5-Flash	text-embedding-small-3	1.5h	∼1.0k	∼1.0k	75.3	53.2	75.3	2.68
MemOS (Li et al., 2025)	Gemini-2.5-Flash	BAAI/bge-m3	4.0h	∼3.1k	∼3.1k	27.5	32.3	27.3	1.09
AI Agent									
Claude Code (Anthropic, 2025)	Sonnet-4.6	*proprietary, unclear trace*	0.0h	∼59.0k	∼59.0k	–	<u>54.0</u>	–	<u>3.77</u>
`camroll-agent` (ours)	Gemini-2.5-Flash	all-MiniLM-L6-v2	1.5h	∼3.2k	∼15k	88.5	**70.5**	83.1	**4.11**

captions and adding image-embedding retrieval baselines (e.g., CLIP or SigLIP top-k images). We view these as important extensions but leave them for future work, as they would require substantial re-engineering of each baseline system. The main takeaway is that modality-preserving memory with targeted pixel inspection is a viable design; the specific gains are an artifact of the full system.

6 Conclusion and Discussion

We introduced `camroll`, a benchmark for camera-roll-grounded question answering, together with `camroll-agent`, a modality-preserving agent for navigating longitudinal visual histories. `camroll-agent` reaches 70.5% MCQ accuracy with ∼15k total query-time tokens (∼3.2k in the answering model's context)—retaining about 92% of the all-images accuracy at roughly $50\times$ fewer total tokens (∼$200\times$ fewer tokens in the answering model's context)—and within-agent ablations show the hierarchy and dedicated retrieval/inspection tools each contribute. These findings support modality-preserving progressive access as a practical design principle: retrieve through compact abstractions while retaining the source modality for targeted verification. The benchmark focuses on annotator-authored, visually grounded queries across 50 users and two capture regimes; its free-form metric lacks human validation or user-level uncertainty estimates. For reproducibility, we ship a fully public core: the YFCC subset (20 users, 1,000 questions) is redistributable today under Creative Commons and reproduces our full pipeline immediately, while the in-house subset (30 users, 1,500 questions) is access-controlled pending legal and consent review. These boundaries motivate owner-centered evaluation, source-stratified analysis, caption-factuality measurement, matched-budget comparisons, and scaling to larger rolls; we are already extending `camroll` to more users via YFCC-100M (Thomee et al., 2016) and to larger per-user histories. Future work should also study learning-based retrieval, joint training, and privacy-preserving user-controlled memory.

AI use statement

We used AI tools in the preparation of this work. Claude Opus 4.8 was used to improve the clarity and readability of author-written text. Claude Code (Opus 4.8) and OpenAI Codex (GPT 5.6 Sol) were used as coding assistants during implementation. These tools were not used to collect data, run experiments, or generate reported results. All AI-assisted text, interpretations, and code were reviewed and verified by the authors, who take responsibility for the final content of this work.

Reproducibility statement

We provide comprehensive implementation details and reproducibility information throughout the paper and appendix. Section 3 describes the `camroll` dataset, including data sources, filtering

criteria, and annotation protocols. Section 4 details the `camroll-agent` method, including the hierarchical memory design (personalized captions and event segmentation), tool definitions, and the exact prompts used for caption generation and event segmentation. Section 5 specifies the experimental setup and evaluation metrics (MCQ accuracy and LLM-as-Judge scoring). The Appendix provides additional ablation studies on memory structure and tool components, model and hyperparameter comparisons across proprietary and open-source backbones, and agent behavior analysis.

REFERENCES

Use apple intelligence in photos on iphone. https://support.apple.com/guide/iphone/use-apple-intelligence-in-photos-iphf7de217f0/ios. Apple Inc., Accessed: 2026-05-06.

Chatgpt. https://chat.openai.com. OpenAI, Accessed: 2026-05-06.

Claude. https://claude.ai. Anthropic, Accessed: 2026-05-06.

Gemini. https://gemini.google.com. Google, Accessed: 2026-05-06.

Copilot + onedrive: Intelligence in every click, inspiration in every memory. https://techcommunity.microsoft.com/blog/onedriveblog/copilot--onedrive-intelligence-in-every-click-inspiration-in-every-memory/4458882. Microsoft, Accessed: 2026-05-06.

Affenstunde. Digital hoarding: Why your phone has 10,000 photos you'll never look at, 2025. URL https://affenstunde.com/2025/07/15/digital-hoarding-why-your-phone-has-10000-photos-youll-never-look-at/. Accessed: 2026-04-26.

Meta AI. Introducing muse spark: Msl's first model, purpose-built to prioritize people. https://ai.meta.com/blog/introducing-muse-spark-msl/, 2026. Accessed: 2026-05-04.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries, 2024.

Anthropic. Claude Code: Agentic coding. https://www.anthropic.com/product/claude-code, 2025. Accessed: 2026.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *ACL 2024*, 2024.

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

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. In *arXiv*, 2026.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 2024.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts llm performance despite perfect retrieval. In *EMNLP*, 2025.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation. In *ICLR*, 2026.

David Fernández-Pérez, Lucía Ros, and José M. Latorre. The role of the personal relevance of images in retrieving autobiographical memories for emotion regulation: A randomized controlled trial study. *Current Psychology*, 43:3523–3537, 2024. doi: 10.1007/s12144-023-04582-5.

Google. Ask photos: Search your memories in google photos. https://blog.google/products-and-platforms/products/photos/ask-button-ask-photos-tips/, 2024. Accessed: 2026-05-04.

Google. Personal intelligence in gemini app with nano banana. https://blog.google/innovation-and-ai/products/gemini-app/personal-intelligence-nano-banana/, 2026. Accessed: 2026-05-04.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *ICML*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. In *arXiv*, 2025.

Lisa Kislinger and Kurt Kotrschal. Hunters and gatherers of pictures: Why photography has become a human universal. *Frontiers in Psychology*, 12, 2021.

Jiahao Nick Li, Zhuohao (Jerry) Zhang, and Jiaju Ma. Omniquery: Enabling question answering on personal memory by augmenting multimodal album data. In *UIST*, 2024.

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. In *arXiv*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. In *ACL*, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arxiv*, 2024.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to me: Long-term personalized referential memory qa. In *arXiv*, 2026.

Mixbook. Survey: The states that phlush away the most memories, 2023. URL https://www.mixbook.com/inspiration/states-that-phlush-away-memories. Accessed: 2026-05-06.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your personalized language and vision assistant. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=mjGy8g3pgi.

Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'chameleon: Personalized vision and language generation. *CVPR*, 2025.

Viet Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal visual memory from explicit and implicit evidence, 2026. URL https://arxiv.org/abs/2605.28806.

Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. Personavlm: Long-term personalized multimodal llms. In *CVPR*, 2026.

PhotoAid. Mobile photography statistics, 2024. URL https://photoaid.com/blog/mobile-photography-statistics/.

Luming Tang, Nataniel Ruiz, Chu Qinghao, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668*, 2023.

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. 2016.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 2002.

Yufei Wang, Zhe Lin, Xiaohui Shen, Radomĭr Měch, Gavin Miller, and Garrison W. Cottrell. Recognizing and curating photo albums via event-specific image importance. In *BMVC*, 2017.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *ICLR*, 2025.

Tianyi Xu, Rong Shan, Junjie Wu, Jiadeng Huang, Teng Wang, Jiachen Zhu, 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 *arXiv*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *arXiv*, 2023.

A Appendix

Table 4: Comparison of base/build model combinations across proprietary and open-source settings. Scores (overall, 10:40 semantic/episodic-weighted judge) and input tokens are computed on the in-house split (30 users, 1,500 free-form questions), judged by GPT-4o on a 0–10 scale. Input tokens are lower than in Table 3 (∼3.2k) because this split contains only in-house rolls. Absolute values are not directly comparable to the full-benchmark (50-user, MCQ+free-form) numbers. The Gemini-2.5-Flash / Gemini-2.5-Flash row is the same run as the Final agent in Table 5.

Base model	Build model	Input	Judge↑
Proprietary *– exposure all data*			
Gemini-2.5-Flash	Gemini-2.5-Flash	2.4k	5.04
Gemini-3.1-Preview-Pro	Gemini-2.5-Flash	16.7k	**5.80**
GPT-4o	Gemini-2.5-Flash	5.9k	3.88
GPT-5.2	Gemini-2.5-Flash	7.6k	<u>5.45</u>
Proprietary *– exposure partial data (build locally)*			
Gemini-2.5-Flash	Qwen3-VL-8B-Instruct	6.5k	3.64
Gemini-3.1-Preview-Pro	Qwen3-VL-8B-Instruct	14.8k	**5.30**
GPT-4o	Qwen3-VL-8B-Instruct	5.4k	3.57
GPT-5.2	Qwen3-VL-8B-Instruct	7.2k	<u>4.99</u>
open-source *– totally local, private*			
Qwen3-VL-8B-Instruct	Qwen3-VL-8B-Instruct	6.0k	2.05
Qwen3-Coder-30B-A3B	Qwen3-VL-8B-Instruct	13.3k	**3.82**
GLM-4.7-Flash†	Qwen3-VL-8B-Instruct	8.7k	<u>2.60</u>

Table 5: Ablation study on memory structure and tool usage, reporting semantic and episodic performance and the overall score ("J" denotes the LLM-as-Judge score; Overall is the 10:40 semantic/episodic-weighted mean). Scores and input tokens are computed on the in-house split (30 users, 1,500 free-form questions). Input tokens are lower than in Table 3 (∼3.2k) because this split contains only in-house rolls. Absolute values are not directly comparable to the full-benchmark numbers. The Final agent row is the same run as the Gemini-2.5-Flash / Gemini-2.5-Flash row in Table 4.

Setting	Input ↓	J-sem↑	J-epi ↑	Overall ↑
Final agent	2.4k	**5.90**	**4.82**	**5.04**
Ablation about memory structure				
Generic caption	4.2k	4.03	4.00	4.01
No Event	3.6k	4.36	3.95	4.03
No Caption	2.7k	3.25	2.04	2.29
Ablation about tools				
No Search	2.7k	2.63	2.94	2.88
No Grep	4.0k	4.25	3.90	3.97
No List	3.7k	4.28	4.04	4.09
No Get	4.3k	4.44	3.99	4.08
No View	3.5k	3.52	3.15	3.22

A.1 Ablations

We conduct ablations on the in-house split (30 users, 1,500 questions), which reflects modern mobile camera rolls. Unlike the cross-system comparison, these ablations preserve the agent configuration and change one memory or tool component at a time. We report free-form judge scores (0–10), which

Table 6: Robustness of free-form evaluation to the judge model. Re-scoring with Gemini-2.5-Flash preserves the system ordering and the main observations obtained with GPT-4o. This comparison measures judge-model sensitivity and is not a substitute for human validation.

Method	MC Recall ↑	MC Acc ↑	FF Recall ↑	GPT-4o Judge ↑	Gemini Judge ↑
Nothing	0.0	30.0	0.0	0.00	0.00
All captions	100.0	63.4	100.0	3.82	4.12
All images	100.0	76.5	100.0	5.01	5.25
Oracle	100.0	86.4	100.0	6.33	6.57
Self-RAG	25.8	46.2	19.8	2.41	2.61
HippoRAG2	50.1	48.5	50.1	2.58	2.63
SimpleMem	57.8	44.6	58.6	1.70	1.81
LightMem	70.3	52.7	70.2	2.44	2.70
Mem0	75.3	53.2	75.3	2.68	2.89
MemOS	27.5	32.3	27.3	1.09	1.22
Claude Code	–	54.0	–	3.77	3.84
`camroll-agent` (ours)	88.5	70.5	83.1	4.11	4.40

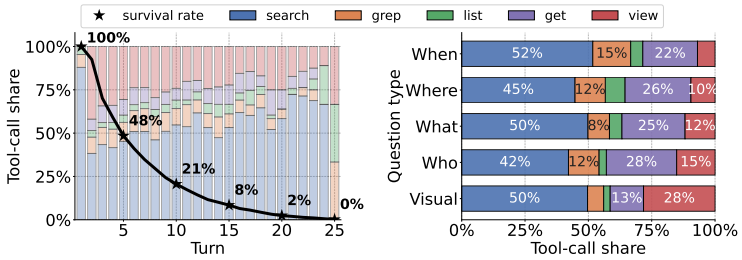


Figure 4: Tool-call distributions across turns and question types.

Error category	%
A. Wrong evidence	14.7
B. Right evidence, passed image	24.7
C. Ran out of steps	10.5
D. Gave up prematurely	21.7
E. Right evidence, flawed answer	17.5
F. Other	10.9
Total	100.0

Table 7: Error analysis on incorrect questions.

capture partial degradations in answer quality and evidence grounding more sensitively than binary MCQ accuracy. Large, consistent drops (e.g., 5.04 → 2.29 without captions, 5.04 → 2.88 without search) persist across independent judges (Table 6).

Memory structure ablation. As shown in Table 5, the full memory design achieves the best performance within this ablation (overall 5.04, the 10:40 semantic/episodic-weighted mean of 5.90 and 4.82). Removing structure degrades results consistently: generic captions drop the overall score to 4.01 (semantic 4.03, episodic 4.00), removing events reduces it to 4.03, and removing captions causes the largest degradation (overall 2.29, episodic 2.04). Thus, captions are the most influential tested memory component under this configuration. However, this ablation measures downstream utility rather than caption factuality; profile-conditioned captions may improve retrieval while also propagating incorrect identity or relationship assumptions.

Tools ablation. The full system reaches overall 5.04 (Tab. 5). Removing search causes the largest drop, while removing grep/list/get/view yields smaller but consistent degradations. Within this configuration, all tools contribute and search has the largest measured effect on performance.

MLLMs. Closed-source models perform better than open-sourced: Gemini-3.1-Preview-Pro achieves the top score (5.80, 5.30; inputs 16.7K / 14.8K across builds), followed by GPT-5.2, Gemini-2.5-Flash, and GPT-4o. While there is a gap, the best open-source model performance is already close to GPT-4o, suggesting a viable alternative for running our agent locally (See Table 4). These results show that the workflow is robust across several proprietary and open-source backbones and supports fully local configurations.

A.2 Agent Behavior Analysis

How agents spend their tool budget. Figure 4 (left) shows the per-turn tool-call distribution, with the survival curve (⋆) marking episodes still active at each turn. Agents open with coarse retrieval (`search`, `grep`, `list`) for broad candidate discovery, then switch to `get`/`view` for inspection;

nearly half of episodes terminate by Turn 5, so most questions resolve in a few rounds. For the harder questions that survive longer, the share of coarse retrieval rises again—agents expand the search space rather than re-inspecting the same evidence—before relying on symbolic retrieval or raw-pixel `view` for a final decision. On the right, *Visual* questions allocate a much larger share to `view`, *when* questions lean on `list`, and *what/who* questions are search-heavy. The benchmark therefore exercises genuinely different tool-use skills across question types, rather than a single retrieval pattern applied uniformly.

When agents fail and why. We assign each failing trajectory to one of six mutually exclusive categories using deterministic rules over the trace (step count, view calls, evidence recall; Table 7). Failures are identified via MCQ accuracy. Most failures stem from agent *decisions* (A–D) rather than the final visual answer: missing or unopened relevant images (A, B), exhausting the budget (C), or stopping too early (D). Only 17.5% (E) are cases where the agent reached the right evidence but answered incorrectly. Category B (unopened relevant images) is not purely a policy failure: the agent answered from a caption, so a wrong answer there may reflect caption-level perception errors made during memory construction. These patterns motivate better search calibration, caption quality, and post-training.

Do we need domain-specific agents? A generalist coding agent can be repurposed for camera roll setting, but its tool inventory imposes a strong inductive bias toward filesystem traversal and byte-level inspection. As shown in Fig. 5, Claude Code lacks a semantic index and therefore alternates between search (e.g., Bash/Glob, 45.3%) and exhaustive visual inspection (Read, 51.9%), leading to inefficient investigation of relevant images and yielding high token usage (59.0k). In contrast, our `camroll-agent` allocates the majority of its budget to a domain-specific semantic retrieval tool (53.6% `search`), requiring far fewer image views (25.2%) (∼3.2k context tokens; ∼15k total including images processed by view, versus ∼59.0k for Claude Code). This mismatch suggests a domain-specific semantic index is an important efficiency lever outside code repositories; since Claude Code also uses a different backbone and tool environment, the comparison is illustrative rather than a controlled test of tool design alone.

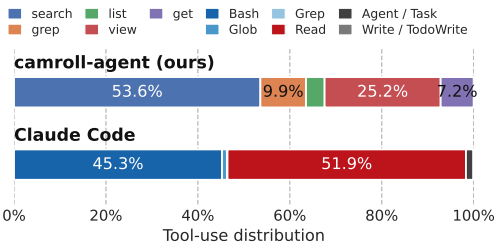


Figure 5: Claude Code vs. `camroll-agent` tool call distributions.

A.3 Experimental Details

Memory store and indices. The database uses two normalized SQLite tables—$\mathcal{I}$ (images and their captions) and $\mathcal{E}$ (events)—linked by a foreign key from $\mathcal{I}$`.event_id` to $\mathcal{E}$`.event_id`. On top we build a BM25 lexical index (SQLite FTS5) for exact matching and verification, and a dense FAISS index ("sentence-transformers/all-MiniLM-L6-v2") for semantic retrieval under paraphrase or abstraction, yielding a hierarchical memory of raw images, image-level captions, and event-level summaries, all traceable via stable hashed identifiers (`img_<h>` and `ev_<h>`). We set the look-back window to 3; the tool budget is at most 25 tool calls, and `view` is capped at 5 calls (with at most 6 images visible at once).

Baseline configuration. For the all-images baseline, we resize each image to a maximum height of 768px to fit Gemini-2.5-Flash's context window and file-upload limit; this downscaling may disadvantage the baseline on fine-grained perception questions, so its accuracy should be read as a strong but not strictly upper-bound reference. Claude Code is run with a budget of $0.5 per question.

Metric details. Evidence recall credits a gold image as recalled if retrieved directly or if any event containing it is retrieved. This event-level crediting is permissive compared to per-image retrieval and is used as a within-system diagnostic. Input tokens count cumulative tokens (reasoning, input, context, tool calls, retrieved results) before the final answer, reported separately from memory-construction cost. Token counts serve as order-of-magnitude estimates across systems due to differences in tokenizers and image-token conventions. Reported scores are point estimates without judge variance or user-level confidence intervals.

Compatibility. All interaction is mediated through these five tools, so the design of `camroll-agent` is model-agnostic: swapping the LLM, the captioner, or the retrieval backend requires only replacing the corresponding component, while the agent loop and the tool interface stay unchanged. This modularity enables the controlled within-agent ablations in Appendix A.1. We expect that jointly fine-tuning the LLM and the tools would yield further gains and leave this for future work.

Computational requirements. For a 1000-image roll, memory construction costs approximately \$0.6–\$5 with the proprietary APIs tested here, or about 1.5 hours on a single 24 GB GPU; after this one-time build each query uses roughly 3.2k input tokens ($\sim$\$0.001–\$0.01). The build is a single pass over the roll—comparable in cost to feeding all images once—so at $\sim$3.2k versus $\sim$750k query-time tokens, the amortized cost of `camroll-agent` drops below repeatedly sending the entire roll within the first few queries per roll. Open-source configurations (Table 4) run fully locally on a 24 GB GPU.

A.4 Broader Impacts

This work studies long-horizon reasoning over personal camera rolls, a setting with potential applications in personalized AI assistants, memory support, and multimedia retrieval. At the same time, personal photo collections contain highly sensitive information, including identities, relationships, locations, and daily activities. Systems with persistent multimodal memory therefore raise important privacy and security concerns, including risks of unauthorized retrieval, profiling, or memorization of personal content. The YFCC subset follows its original Creative Commons licenses, whereas broader redistribution of the consented in-house subset remains under legal and consent review. We also note two further limitations. First, personal photos frequently contain third parties (e.g., friends, family, or bystanders) who did not directly consent to inclusion; we treat this as an important limitation and restrict redistribution of the in-house subset accordingly. Second, our memory-construction pipeline sends personal images to proprietary captioning APIs, which is itself a privacy consideration; the open-source configurations in Table 4 offer a fully local alternative that avoids exposing images to third-party services. Future deployments should prioritize user consent, controllable memory management, secure storage, and privacy-preserving mechanisms. We hope this work encourages further research on safe and transparent personalized multimodal memory systems.

A.5 Data Statistics

Demographics and coverage. Fig. 2 summarizes the geographic and temporal footprint of the `camroll` dataset. Together, the two subsets span 24 years of personal photo-taking (2002–2026) across five continents and 25+ countries. The right panel (Fig. 2b) shows that the two subsets capture two distinct patterns: YFCC peaks in 2006–2010 with the global popularity of dedicated digital cameras, while the in-house subset accelerates sharply from 2023 onward, reflecting the dense, low-curation regime of contemporary smartphones. Per active user, the smartphone era is roughly 1.6$\times$ denser than the digital-camera era: mobile users accumulate on average $\sim$17 photos/month, versus $\sim$11/month for YFCC100M users. Thus, the two subsets have comparable absolute size (15,869 vs. 15,607 images) but substantially different capture eras, devices, resolutions, and temporal density profiles. This diversity broadens the benchmark, but it also creates distinct task regimes. Our aggregate results characterize the combined benchmark and should not be interpreted as separate estimates for contemporary smartphone and historical YFCC camera rolls.

Availability. The two-source design spans two capture regimes—the digital-camera era and the smartphone era—and lets us ship a fully redistributable public core alongside an access-controlled extension: the YFCC core (20 users, 1000 questions) is distributable under the original Creative Commons licenses and runnable by any researcher today, while the in-house extension (30 users, 1500 questions), collected with permission for research use, is held to a higher redistribution bar pending legal and consent review. This mirrors personal-data benchmarks, where the most realistic contemporary data is precisely what cannot be released without care.

Question analysis. We analyze the questions set to understand its linguistic structure and grounding requirements. Episodic questions are roughly 2$\times$ longer than semantic ones (15.9 vs. 7.3 words on average), reflecting the additional contextual information needed to specify time, place, or events. Beyond length, we observe clear differences in question formulation: semantic questions are

dominated by *what*-style queries (57%), whereas episodic questions are more diverse and heavily shaped by temporal and prepositional phrasing (e.g., "on", "in", "after", "during"). About 46.2% of questions can be answered from a single image, while 32.2% require reasoning across multiple images and 20.0% require whole-roll context. Operationally, we label a question *multi-image* when it needs a small, bounded set of images and *whole-roll* when answering requires aggregating or scanning across the roll as a whole—for example, questions about the most frequent or first/last occurrence of something, or establishing absence (e.g., "Have I ever been to Italy?"). This shows that more than half of the questions go beyond single-image VQA and require cross-image or sequence-level reasoning. In addition, 23.8% of questions involve fine-grained perceptual understanding (e.g., counting, OCR, or attribute-level details). Finally, `camroll` is strongly first-person centered (88.4% explicit "I/my/me" usage) and temporally grounded (62.4% contain explicit time or event references), reinforcing its nature as a personal, longitudinal memory benchmark rather than standard visual question answering. These linguistic properties describe the annotator-authored questions in the benchmark; they should not be interpreted as evidence that owners would ask questions with the same distribution.

Answers analysis. Gold answers are typically short but rarely single-word: the median answer length is 2 tokens, and 72.9% are multi-word phrases (mean 2.86, max 15). This reflects that questions about personal life often require precise answers (e.g., a specific outfit, a place name, or a duration), rather than a single object label or overly long descriptions (e.g., image captioning). Episodic answers are slightly longer than semantic ones (mean 2.95 vs. 2.51 tokens), consistent with the need to disambiguate among similar past events. Distractors are written by the same annotator with knowledge of the user's roll, and 89.7% are length-matched to the gold answer to within two tokens, so the format does not leak the correct option through surface form. Most importantly, gold answers are *personal*: Of the 2,084 distinct content tokens (length ≥ 4) that appear in gold answers, 66.9% appear in only one user's answers, and 88.2% of unique answer bigrams appear in only a single user's roll. Of the 1,875 distinct full answer strings, only 9.8% are reused across two or more users, and the most frequently repeated answers are exactly those with weak personalization signal (*white*, *student*, *yellow*, *red*). This indicates that solving `camroll` requires retrieving content from the target user's own album rather than relying on common visual concepts shared across users.

Table 8: Evidence coverage across dataset subsets and memory types.

Subset	Type	% with evidence	n
In-house	semantic	0%	0 / 300
In-house	episodic	97.4%	1,169 / 1,200
YFCC	semantic	0%	0 / 200
YFCC	episodic	98.9%	791 / 800
Combined	all	78.4%	1,960 / 2,500
Combined	episodic only	98.0%	1,960 / 2,000

Error categorization. We classify each *incorrectly*-answered question (LLM-judge score $= 0/10$) into one of six mutually-exclusive categories, evaluated in the order listed.[1] Let s denote the number of agent actions in the trace, v the number of `view` calls, and $\rho \in [0, 1]$ the RECALL_IMG_OR_EVENT signal: the fraction of ground-truth evidence images that appeared (by stem or via a containing event name) in *any* tool result of the trace. Let $\mathcal{G}$ be the set of ground-truth evidence image stems and $\mathcal{V}$ the set of stems the agent actually opened with `view`.

The categories below use the same labels (A–F) as Table 7; we list them in the top-to-bottom order in which the rules are evaluated rather than in alphabetical order.

(C) Ran out of steps / budget. The agent exhausted its action budget: either `stopped_reason = max_steps`, or it used at least 20 of the 25 allowed actions (a near-budget threshold that also captures traces terminating just short of the hard cap), or it hit the per-trace `view` cap of 5 calls. These traces ended because of a hard limit, not because the agent decided it was done.

[1] Categories are evaluated top-to-bottom; the first matching rule wins, so each question lands in exactly one bucket.

Table 9: Distribution of licenses in the initially downloaded 15,927-image pool from 20 YFCC-100M users, before final file validation and dataset assembly.

License	Count	%
CC BY-NC-SA 2.0	5,438	34.1
CC BY 2.0	4,945	31.0
CC BY-NC 2.0	4,028	25.3
CC BY-SA 2.0	1,516	9.5
Total	**15,927**	**100.0**

Table 10: Composition of `camroll`. The two subsets are complementary: the in-house subset captures contemporary smartphone behavior at full resolution with rich participant-authored event labels, while YFCC contributes longer per-user spans, real EXIF/GPS metadata, and a publicly redistributable license at lower resolution. *Encoded in the filename (`YYYY-MM-DD HHMMSS.jpg`); †encoded in the YFCC100M `datetaken` metadata field; ‡YFCC also contains a smaller fraction of early-smartphone captures (e.g., iPhone 4).

Property	In-house	YFCC	Total
Users & images			
Number of users	30	20	50
Total images	15,869	15,607	31,476
Images per user (mean / median)	529 / 533	780 / 808	630 / 558
Images per user (min / max)	225 / 707	505 / 979	225 / 979
Temporal coverage			
Capture date range	2019 – 2026	2002 – 2014	2002 – 2026
Per-user span, years (mean)	2.7	6.6	4.3
Per-user span, years (min / max)	1.6 / 3.8	3.0 / 9.4	1.6 / 9.4
Images with second-precision timestamp	98%*	100%†	99%
Image properties			
Resolution, mean (W × H)	3188 × 3500	592 × 514	–
Resolution, median (W × H)	3072 × 4000	500 × 480	–
Megapixels (mean / median)	11.7 / 12.2	0.43 / 0.19	–
Orientation (landscape / portrait / square)	33% / 64% / 3%	60% / 23% / 17%	46% / 44% / 10%
File size, median	3.0 MB	61 KB	–
Total dataset size	64 GB	2.3 GB	66 GB
Capture metadata			
Capture device	smartphone	digital camera‡	mixed
Geo-tagged images	–	75%	37%
Profile photo	✓	✓	✓
Profile biography (text)	–	13 / 20 users	–
Annotations			
User-labeled events	841 (28/user)	–	841
Avg. event size (# images)	9.7	–	–
QA pairs (semantic + episodic)	300 + 1,200	200 + 800	500 + 2,000
QA pairs, total	1,500	1,000	**2,500**
Episodic Qs grounded to image(s)	97%	99%	98%

(D) Gave up prematurely. The agent voluntarily stopped (`stopped_reason = ok`) after at most `PREMATURE_STEPS` $= 2$ tool calls. The agent answered with very little exploration.

(A) Wrong evidence. Ground-truth evidence exists but the trace failed to retrieve it ($\rho < 1$). The retrieval pipeline did not surface all of the right images / events.

(B) Right evidence, passed image (never looked). All gold evidence was retrieved ($\rho = 1$) but the agent never invoked `view` on any of the gold images ($\mathcal{G} \cap \mathcal{V} = \emptyset$). The agent answered over-confidently on a search-result snippet without inspecting the image itself.

Table 11: Question-type schema. Each question is assigned exactly one label by an LLM classifier (GEMINI-2.5-FLASH), based on the shape of the gold answer rather than on the question's surface form. Tie-breaking priority: VISUAL > WHEN > WHERE > WHO > WHAT. Counts are reported over the in-house split (1,500 questions); the YFCC split is not included in this particular analysis.

Label	Answer is…	Example question	n
What	an object or action	"What did I eat for breakfast?"	611
Where	a place, venue, or location	"Where did I eat dinner the day before going to the museum?"	173
When	a date, duration, or temporal order	"When did I last see my grandfather?"	123
Who	a person or group	"Who came to my birthday party in 2024?"	75
Visual	a visual attribute or exact count from a photo (color, count, written text, breed, fine-grained appearance)	"How many balloons were in the photo?"	518
		Total	1,500

Table 12: Condition-type schema. Conditions are the constraints in the question that scope the search (which photo / event to look at), separately from the question type (what attribute to extract). A single question may carry zero, one, or several conditions; counts therefore sum to more than 1,500. The condition vocabulary is intentionally distinct from the question-type vocabulary so the two slots are not conflated. Counts are reported over the same in-house split (1,500 questions) as Table 11.

Label	Triggered by…	Example phrase in the question	n
Situation	an episode, activity, or object that scopes the search	"the day I visited the museum", "during my road trip", "at my friend's wedding"	1,182
Location	a place, venue, or location reference	"in Beijing", "at the zoo", "in the kitchen"	621
Time	a date, year, time of day, or temporal anchor	"in 2024", "the day after", "on Christmas"	778
Person	a specific person whose presence scopes the search	"with my mom", "when I was with Lin"	423
(none)	question imposes no scoping constraint, e.g. "What did I eat for breakfast?"		235

(E) Right evidence, flawed answer (looked, still wrong). All gold evidence was retrieved ($\rho = 1$) *and* the agent explicitly opened at least one gold image with `view` ($\mathcal{G} \cap \mathcal{V} \neq \emptyset$), yet still produced a wrong answer. This is a genuine perception-detail or reasoning failure on inspected content.

(F) Other. The question carries no ground-truth evidence list, or the evidence signal is unavailable, so the (A)/(B)/(E) distinction does not apply (e.g., semantic questions).

A.6 PROMPTS

The captioning prompt below permits cautious first-person and relationship language to improve personal retrieval. Such phrases remain model inferences rather than verified annotations. In particular, labels such as "my friend" can be incorrect and can persist in later event summaries; evaluating and reducing this form of memory error is left for future work.

Image Captioning and Event Segmentation Prompt

You are maintaining long-term structured memory for one user's personal photo library.
You will see two images: 1. The user's profile photo. 2. The current image that must be processed.

Important: - Use the first image only as identity/reference image. This image will tell you how the user looks like. - You should write the caption and event decision from the perspective of the user in the first image. - The second image as the only source for “image-caption”, “operation”, and event reasoning. - Never describe the first image in “image-caption”. - The album is processed strictly in chronological order from oldest to newest. - Only update the most recent event row if the current image clearly belongs to it.
Event definition: - An event should be an episodic memory unit, such as a trip, outing, meal, hangout, celebration, class activity, or other coherent real-world episode. - An event can span multiple consecutive photos with different dates, locations, subjects, poses, or close-up details, as long as they still belong to the same broader episode (e.g., a road trip, hangout with friends, selfies, etc.). - Event names should summarize the broader episode, not the most eye-catching object in a single frame.
Album description: `\{library-description\}`
Current image metadata: - date: `\{current-date\}`
Recent image table rows (up to the last k images):

```
json.dumps(recent-payload, indent=2, ensure_ascii=False)
```

Latest event summary:

```
json.dumps(latest_event_payload, indent=2, ensure_ascii=False)
```

Tasks: 1. Write a detailed personalized caption for image 2 (the current album image) as if the person in image 1 is describing their own photo in first person.
2. Choose exactly one operation for the event table: - ADD: create a new event row - UPDATE: update the latest event row - NO OP: do not modify the event table
3. If you choose ADD or UPDATE, return a full event row with fields: - event name - description - date - images
Rules: - The image caption must always be present. - Use first-person wording when natural. - The image caption must describe image 2 only, not the profile/reference image. - Be careful about person identity; the person in the first image is the user, and the person in the second image might be different. - If identity is unknown, use neutral terms like "a man", "a woman", or inferred relations like "my friend". - Make the image caption specific and detailed. - Mention visible content such as setting, people, objects, activity, and atmosphere. - Write as a personal memory grounded only in image 2. - Do not hallucinate precise facts not supported by the image.
Event update rules: - For ADD or UPDATE, include current image path in images. - Prefer UPDATE if the image is part of the same ongoing episode. - Prefer ADD only when a clear new event boundary exists. - Keep event descriptions under 300 words and concise. - Event names should reflect broad activities (e.g. "Trip to Chengdu", "Coffee Shop Hangout"). - Prefer NO-OP if nothing meaningful changes.
Output format (VALID JSON ONLY, no markdown):

```
{
  "operation": "ADD" | "UPDATE" | "NO_OP",
  "image_caption": "string",
  "event": {
    "event_name": "string",
    "description": "string",
    "date": "string",
    "images": ["full image path", "..."]
  } | null
}
```