# Personal Visual Context Learning in Large Multimodal Models

**Zihui Xue** **Ami Baid** **Sangho Kim** **Mi Luo** **Kristen Grauman**
The University of Texas at Austin

## Abstract

As wearable devices like smart glasses integrate Large Multimodal Models (LMMs) into the continuous first-person visual streams of individual users, the evolution of these models into true personal assistants hinges on visual personalization: the ability to reason over visual information unique to the wearer. We formalize this capability as Personal Visual Context Learning (Personal VCL), the prompt-time capability of using user-specific visual context to resolve personalized queries. To systematically evaluate this, we present Personal-VCL-Bench, a comprehensive benchmark capturing the personal visual world across persons, objects, and behaviors. Our analysis of frontier LMMs identifies a profound context utilization gap, revealing that the mechanisms for leveraging visual evidence, as well as aggregating multiple visual observations, remain critically understudied. Motivated by these findings, we propose the Agentic Context Bank, a strong inference-time baseline that structures a user's visual context into a self-refining memory bank and employs query-adaptive evidence selection. Our baseline approach consistently improves over standard context prompting regimes across tasks and evaluated backbones, demonstrating a practical path towards future personalized LMMs.[1]

## 1 Introduction

The immense capabilities of Large Multimodal Models (LMMs) [56, 26, 48, 7, 72] are undeniable. Looking ahead, wearable technologies like smart glasses [20, 49, 35] will embed these models into the continuous first-person visual stream of an individual user. This continuous egocentric perception [29, 30, 13, 34, 78] is the critical catalyst for turning generalized AI into dedicated personal assistants. We envision a future as depicted in Fig. 1. Through daily observation, the model constructs an entirely unique visual profile about the user: it maps their social circle, identifies personal items, and knows exactly how they like their fried eggs cooked. Equipped with this visual memory, AI assistants can answer personalized queries and provide active guidance for the user, such as pinpointing where they left their water bottle or alerting if the egg is getting overcooked compared to their usual standard.

This vision fundamentally rests on the broader challenge of model personalization [82]. In the text domain, the established mechanism for personalizing LLMs [63, 42, 53, 62, 64] relies on retrieving relevant written information about the user and prepending it to the prompt to guide generation. In this paper, we explore the direct visual analog of this paradigm. As illustrated in Fig. 1, a user's visual history consists of the long, continuous egocentric stream captured over time, from which relevant images or clips are drawn to form a *personal visual context*. When a user poses a specific question, this personal context is supplied directly alongside the query. A critical question then emerges: can today's LMMs effectively reason over this personal visual evidence to resolve user-specific queries? We formalize this capability as Personal Visual Context Learning (Personal VCL).

Our formulation centers around *context utilization*: the ability of an LMM to reason over visual context once it has been supplied in the prompt. This focus distinguishes Personal VCL from

[1] Project webpage: https://vision.cs.utexas.edu/projects/PersonalVCL.

Preprint.

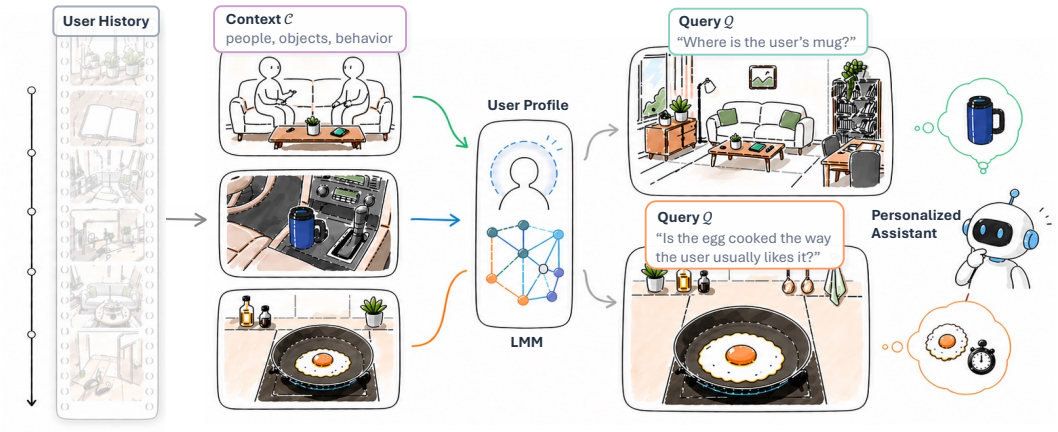


Figure 1: Personal Visual Context Learning. Continuous egocentric capture from wearable devices presents a user's unique visual history, supplying personal context absent from standard model pre-training. Personal VCL investigates whether an LMM can effectively leverage this context to resolve user-specific queries, ranging from locating a personal object to comparing a current action against the user's past behaviors.

three neighboring lines of work. First, existing LMM personalization works enroll static user-specific concepts (like a specific pet or face) from a few reference images [12] via a dedicated training [1, 54, 2, 60, 59, 3, 55, 4, 67, 39] or storage mechanism [31, 65, 6, 32, 83], but cannot address inference-time reasoning over complex, dynamic personal context that unfolds across visual observations. Second, while visual in-context learning [41, 24, 17, 85, 79, 38, 19] shares our inference-time setting, its use of visual demonstrations serves to elicit task formatting from pre-trained knowledge, not to supply novel, private knowledge about a particular user. Finally, long-form [23, 76, 46, 84] and egocentric [29, 21, 37, 14, 50, 78, 10] video question answering (VQA) target *context acquisition*: how to effectively search a long visual history to identify query-relevant evidence. Personal VCL targets the orthogonal subsequent stage: assuming relevant context has been mined from history, it investigates how to best leverage this visual context for personalized queries.

The proposed Personal VCL therefore defines a new capability axis for LMMs: prompt-time reasoning over private, fine-grained, and temporally accumulated visual evidence. The fundamental challenge is that the relevant knowledge is neither part of the model's (pre-trained) world knowledge nor contained entirely in the query. It is carried by the user's personal visual context (mined from the broader visual history): the faces they know, the objects they own, the routines they repeat, and the small deviations that matter only relative to their own past. Advancing this capability could enable personal assistants that reason against an individual's own history, from detecting deviations during physical rehabilitation to assessing skill progression in sports, cooking, or crafts.

To chart a course along this path, we introduce Personal-VCL-Bench, a benchmark organized around the three core axes of a user's visual world: persons, objects, and behaviors. It culminates in EgoWearer identification, a capstone challenge where the model must recognize the user behind a new egocentric video by aggregating subtle, persistent cues from their prior history: who they interact with, what they own, and how they navigate routine activities. Benchmarking frontier LMMs, we reveal a substantial Personal VCL gap: raw visual context is not reliably exploited, and scaling up the volume of context often fails to improve performance.

Motivated by this diagnosis, we propose the Agentic Context Bank as a strong inference-time baseline for Personal VCL. Our framework transforms raw visual context into a structured, evidence-linked memory bank and employs query-adaptive selection to inspect only the necessary visual evidence for a given query. Its consistent gains over standard language- and visual-context prompting, across tasks and LMM backbones, underscore context utilization as the central pillar of Personal VCL. Together, our formulation, benchmark, diagnosis, and baseline approach establish Personal VCL as a concrete capability for building LMMs that can adapt to individuals through visual experience—essential for the next frontier in wearable computing.

## 2 Related Work

**Personalization.** The trajectory of model personalization begins in the LLM domain, where the user is represented as a body of text [82]. Retrieval-based approaches [63, 42, 53, 62, 64] leave weights frozen and pull from the user's written profile on demand, while parameter-based approaches [69, 70, 68] absorb that profile into per-user adapters. The visual counterpart to this evolution is concept-centric: a user-specific entity, such as a pet, a face, or an owned object, is enrolled from a few reference images and bound to a dedicated representation, learned per concept [12, 1, 54, 2, 60, 59, 3, 55, 4, 67, 39], stored in a retrieval database [31], or matched in frozen feature space [65, 6, 32]; recent work further extends the same enrollment paradigm to streaming video [83]. Personal VCL moves beyond this concept-enrollment view and investigates personalization as an act of inference-time reasoning over the broader, dynamic visual context of a user's life.

**Context learning.** The role of the prompt has expanded beyond simple task specification, increasingly serving as a temporary knowledge base for LLMs during inference [44, 25]. A capability gap recently identified by CL-bench [18] separates two roles the prompt context can play: it can demonstrate a known task execution pattern, or supply novel knowledge required to answer, with the second proving far harder. The first role is the one that the dominant in-context learning literature has explored, originating in LLMs [9, 52, 16] and carried into the multimodal setting through image [85, 38, 19] and video demonstrations [79, 41, 24, 17]. In those setups, prompt examples act as structural guides that teach the model an input-output mapping rule for tasks reliant on pre-trained knowledge, such as assigning a class label to an image or an action category to a video. In contrast, Personal VCL positions visual context as a private knowledge base. Because the required personal information is inherently absent from any LMM's pre-training, the model must actively reason over the provided visual evidence to discover the novel facts needed to formulate an accurate response.

**Long-form and egocentric video understanding.** Personal VCL builds on two closely related lines of video research. In long-form VQA [23, 76, 46, 84], the central challenge is context *acquisition*: finding the small set of query-relevant moments from a much longer video. This has driven the development of temporal localization [43, 81], planner-observer agents [61, 45, 75, 5, 40, 73], and memory-augmented stores [66, 77, 51, 8, 74, 36, 11, 15, 80] to construct candidate visual context. Meanwhile, egocentric VQA brings these challenges to first-person capture, featuring tasks like instance search [29, 21], grounded and task-oriented QA [37, 14], and multi-day comprehension [50, 78, 10]. Beyond QA, prior studies on wearer identity and privacy [33, 71, 47] establish that first-person footage inherently encodes the camera wearer's identity, an insight we adopt to formulate EgoWearer identification in Personal-VCL-Bench. Ultimately, while both VQA domains emphasize acquiring long-horizon visual evidence, Personal VCL isolates the critical subsequent step of context *utilization*: given context drawn from the user's history, the model must use it to respond to a user-specific query—even when the response requires indirect reasoning.

## 3 Personal Visual Context Learning

This section formalizes the problem of Personal VCL (Sec. 3.1) and introduces Personal-VCL-Bench (Sec. 3.2). We leverage this benchmark to conduct an in-depth empirical investigation into frontier LMM performance (Sec. 3.3), and derive a strong baseline approach, the Agentic Context Bank, from these findings (Sec. 3.4).

### 3.1 Problem Formulation

Personal VCL assesses whether a model can leverage personal visual history as context to answer user-specific visual queries. Formally, each instance is a pair $(\mathcal{C}, \mathcal{Q})$. The context $\mathcal{C} = (\mathcal{L}_c, \mathcal{V}_c)$ pairs a language declaration $\mathcal{L}_c$ that names what the context shows with the corresponding visual material $\mathcal{V}_c$ it refers to. Because personal context accumulates over time, we consider $\mathcal{V}_c$ as a collection of reference observations from the user's history, such as several images of the same person or object, or several clips of the same routine. The query $\mathcal{Q} = (\mathcal{L}_q, \mathcal{V}_q)$ is structured analogously: $\mathcal{L}_q$ poses a question about the user in relation to the image or video $\mathcal{V}_q$. This language query $\mathcal{L}_q$ may correspond to an explicit user request, or to an internal decision point generated by a broader assistant system (e.g., a proactive agent evaluating whether a user's current physical execution deviates from their


Figure 2: We propose Personal-VCL-Bench to evaluate how LMMs use personal visual context across three axes: persons, objects, and behavior. Each panel shows a representative context–query example. The EgoWearer Identification task serves as the benchmark capstone, where the model must decide whether a query clip belongs to the same camera wearer by connecting indirect cues across axes, such as hands, the phone, and the swiping motion in this example.

historical baseline before triggering an automated correction). A generic pretrained LMM $f_\theta$, frozen at inference time, is tasked with responding to the query $\mathcal{Q}$ based on user context $\mathcal{C}$.

### 3.2 Personal-VCL-Bench

What constitutes our personal visual world? It is fundamentally shaped by the specific people we encounter, the unique objects we interact with, and the distinct activities we perform. Grounded in this philosophy, we introduce Personal-VCL-Bench, structured precisely along these three fundamental axes of personal visual knowledge: persons, objects, and behavior, with EgoWearer identification serving as an integrative task over all three.

We instantiate these axes using three egocentric datasets (EgoLife [78], Ego4D [29], and Captain-Cook4D [58]) that capture the day-to-day perspectives of individual users across multiple scenes. We manually design the task semantics, formulate user-specific queries, and curate the visual contexts needed to answer them. This yields 2,255 clean context-query instances spanning 7 tasks; each query is paired with 4.25 supplied context images/clips on average. Fig. 2 illustrates concrete benchmark instances across all axes, showing how personal context is paired with a user-specific query. See Supp. A for full benchmark details.

**Persons.** We posit that a personal assistant must reliably navigate the immediate social circle of a user. To evaluate this capability, we instantiate our persons axis on EgoLife [78], a multi-day egocentric recording of six co-living participants, via two progressive tasks. The first is single identity recognition, acting as an essential proof of concept for matching a target person using a provided reference gallery. The second task, derived from the RelationQA set of EgoLifeQA benchmark [78], requires the model to resolve multiple people's identities from context and then map those specific identities onto the query video to successfully interpret the social dynamics occurring between them.

**Objects.** The true utility of a visual assistant lies not in recognizing a generic category like a bottle, but in identifying the exact bottle owned by the user. Queries along this axis fundamentally demand instance-level rather than category-level reasoning. We ground this axis by repurposing the Ego4D-VQ benchmark [29] into two complementary tasks: object identity recognition, where the model decides whether the context object appears in a query video, and personalized object detection, where the model localizes the context object instance in a query frame with a bounding box.

Table 1: Diagnosing frontier LMMs on Personal-VCL-Bench across five context regimes (per-task accuracy, %). **Bold** = best within a model's five rows. $k$=1 denotes using a single context item, while $k$=$k_{\max}$ denotes using all available context items for each query. In visual-context rows, cell shading indicates whether visual context improves, hurts, or ties against the matched language-context row. In visual-ctx ($k_{\max}$) cells, superscript markers indicate whether increasing visual context from $k$=1 to $k_{\max}$ improves, drops, or ties. Taken together, the results reveal that current LMMs struggle to capitalize on the primary assets of Personal VCL: utilization of raw visual evidence and integration of multiple context observations. Full results are in Tab. 5 of Supp. B.

| Context regime | Persons | | Objects | | Behavior | | EgoWearer |
|---|---|---|---|---|---|---|---|
| | ID | Rel | ID | Det | Err | QA | ID |
| Random guess | 50.00 | 25.00 | 50.00 | – | 50.00 | 25.00 | 50.00 |
| Qwen3-VL-8B [7] | | | | | | | |
| No-context | 50.50 | 36.00 | 49.06 | 68.12 | 49.46 | 19.39 | 49.83 |
| Language-ctx ($k$=1) | 51.75 | 29.60 | 65.59 | 76.09 | **57.57** | **24.49** | 52.57 |
| Language-ctx ($k_{\max}$) | 62.75 | **48.80** | 67.25 | **77.54** | 54.86 | 22.45 | **54.53** |
| Visual-ctx ($k$=1) | 62.75 | 46.40 | 70.99 | **77.54** | **57.57** | 21.43 | 50.00 |
| Visual-ctx ($k_{\max}$) | **85.25**$^{\uparrow}$ | 46.40$^{\sim}$ | **73.67**$^{\uparrow}$ | 76.81$^{\sim}$ | 55.14$^{\downarrow}$ | 22.45$^{\uparrow}$ | 50.00$^{\sim}$ |
| Gemma-4-31B [22] | | | | | | | |
| No-context | 49.25 | 22.40 | 51.08 | 57.97 | 52.43 | 29.59 | 50.13 |
| Language-ctx ($k$=1) | 57.50 | 20.80 | 78.33 | **72.46** | 54.86 | 33.67 | 54.52 |
| Language-ctx ($k_{\max}$) | **78.00** | **49.60** | **78.75** | 71.74 | 58.11 | **34.69** | 52.99 |
| Visual-ctx ($k$=1) | 65.50 | 30.40 | 73.26 | 63.77 | **59.46** | 32.65 | **56.66** |
| Visual-ctx ($k_{\max}$) | 75.75$^{\uparrow}$ | 46.40$^{\uparrow}$ | 75.34$^{\uparrow}$ | 65.22$^{\uparrow}$ | **59.46**$^{\sim}$ | 27.55$^{\downarrow}$ | 55.36$^{\downarrow}$ |
| Gemini-3-Flash [27] | | | | | | | |
| No-context | 55.75 | 42.40 | 49.32 | 64.49 | 57.84 | 37.76 | 49.47 |
| Language-ctx ($k$=1) | 69.25 | 37.60 | 80.55 | 79.71 | 55.41 | 45.92 | 53.99 |
| Language-ctx ($k_{\max}$) | 81.75 | **59.20** | 82.35 | 78.99 | 54.59 | 48.98 | 55.26 |
| Visual-ctx ($k$=1) | 67.50 | 44.00 | 82.21 | **85.51** | **68.92** | 47.96 | 51.99 |
| Visual-ctx ($k_{\max}$) | **88.00**$^{\uparrow}$ | 54.40$^{\uparrow}$ | **82.47**$^{\sim}$ | 82.61$^{\downarrow}$ | 67.03$^{\downarrow}$ | **61.22**$^{\uparrow}$ | **55.74**$^{\uparrow}$ |

**Behavior.** Human routines are highly individualized, characterized by distinct execution patterns, tempos, and degrees of dexterity. The objective along this axis is to evaluate an action not against a universal standard, but against the user's own baseline. For example, a wearable assistant might track changes in a patient's dexterity during stroke rehabilitation, or compare a cook's current knife technique against their usual safe motion. To achieve this, we adapt CaptainCook4D [58] to formulate two tasks. Given a few reference videos of a user's standard execution as context, a model is tasked with analyzing a new query video of the same user to either identify procedural mistakes or pinpoint fine-grained behavioral consistencies and deviations.

**EgoWearer identification.** A personal assistant is fundamentally anchored to the identity of its user; its memory and actions are only valid if it can maintain continuity over who is wearing the device. We propose EgoWearer identification to test this memory-binding capability: given visual context from one user's history, the model needs to decide whether a new egocentric clip belongs to the same wearer. Crucially, since the first-person viewpoint inherently conceals the wearer, the model cannot rely on direct observation. It must instead construct a coherent profile of the "unseen" user by piecing together indirect visual evidence, such as the objects they repeatedly use, or the distinctive ways they perform routine actions. By demanding this complex deductive reasoning, EgoWearer identification stands as the capstone of Personal-VCL-Bench, providing a stress test for fine-grained personalization and a critical prerequisite for continuous authentication in real-world assistants.

### 3.3 Diagnosing the Personal VCL Gap

To systematically dissect the context utilization problem, we structure our investigation around two core questions: how should personal visual context be represented, and how effectively do models integrate multiple pieces of evidence? For the first question, we evaluate a visual-context regime that provides the reference material $\mathcal{C}$ as raw pixels against a language-context regime that provides

**Algorithm 1. Agentic Context Bank**

**Input:** visual context $\mathcal{V}_c = \{v_1, \ldots, v_k\}$, query $\mathcal{Q} = (\mathcal{L}_q, \mathcal{V}_q)$, LMM $f_\theta$
**Output:** answer $A$

// Stage I: Structured Bank Construction (query-agnostic; Sec. 3.4.1)
$\mathcal{B} \leftarrow \emptyset$
**for** $v_i \in \mathcal{V}_c$ **do**
  $\mathcal{M}_i \leftarrow \textsc{Extract}(v_i)$ // candidate memory entries $(\tau, d, e)$
  **foreach** $(\tau, d, e) \in \mathcal{M}_i$ **do**
    $op \leftarrow \textsc{Merge}(\mathcal{B}, (\tau, d, e))$ // $op \in \{\textsc{add}, \textsc{confirm}, \textsc{revise}, \textsc{retract}\}$
    apply $op$ to $\mathcal{B}$

// Stage II: Query-Adaptive Evidence Selection (query-specific; Sec. 3.4.2)
$T_\mathcal{B} \leftarrow \textsc{TextView}(\mathcal{B})$ // memory descriptors with entry IDs
$y \leftarrow f_\theta(T_\mathcal{B}, \mathcal{L}_q, \mathcal{V}_q)$ // Call 1: text triage
**if** $y = (\text{request}, \mathcal{I})$ **then**
  $H_\mathcal{I} \leftarrow \textsc{HybridView}(\mathcal{B}, \mathcal{I})$ // inline evidence only for selected IDs $\mathcal{I}$
  $A \leftarrow f_\theta(H_\mathcal{I}, \mathcal{L}_q, \mathcal{V}_q)$ // Call 2: selective visual verification
**else**
  $A \leftarrow y.\text{answer}$
**return** $A$

an LMM-generated textual description of that visual content, allowing us to assess whether native visual data offers advantages over textual proxies. A no-context baseline is also established using only the query $\mathcal{Q}$ to quantify the model's reliance on generic pre-trained knowledge. To address the second question, we compare model performance when provided with a single context ($k$=1) versus all available context items ($k$=$k_{\max}$) across both context regimes. We evaluate seven frontier LMMs on Personal-VCL-Bench. Tab. 1 reports a representative subset, with the full results provided in Supp. B. Two key observations emerge:

**The modality paradox.** Human cognition effortlessly processes raw visual snapshots; we recall the exact appearance of a familiar face or a personal object without needing to first translate those memories into words. However, our results indicate that current LMMs struggle to natively emulate this capability. In a great number of our evaluated cases (red/gray cells in Tab. 1), providing the model with a language context, which is inherently a lossy semantic compression of the full visual signal, yields performance that is on par with or even superior to using the raw pixels. This paradox reveals a substantial deficiency in VCL compared to its textual counterpart. We hypothesize that this limitation stems from the foundational training paradigms of frontier LMMs, which remain overwhelmingly optimized for language processing rather than genuine, native visual reasoning.

**The scaling paradox.** A fundamental assumption of in-context learning is that access to more evidence yields better predictions. Yet, in Personal VCL, LMMs contradict this expectation. When comparing a single context item against all available context (1-vs-all; superscript arrows in Tab. 1), the addition of relevant historical data frequently fails to improve accuracy. This reveals a critical inability of current models to synthesize multiple visual observations and extract consistent personal patterns. Since real-world visual histories grow continuously over time, overcoming this scaling bottleneck is imperative for the development of robust personal agents.

### 3.4 Agentic Context Bank

The diagnostic results suggest that Personal VCL is far from being solved by simply appending and concatenating visual context into a prompt. Instead, they point to a key context utilization problem: how should the supplied visual context be best organized and presented to the LMM? Driven by the two observations above, we introduce the Agentic Context Bank, a strong model-agnostic inference-time baseline for Personal VCL. The framework operates in two stages. Stage I constructs a structured bank by converting disjoint context items into a coherent, self-refining memory (Sec. 3.4.1). Stage II applies query-adaptive evidence selection. At inference time, the model first surveys a lightweight text view of the bank, choosing to load the actual visual evidence only for the entries required to resolve the active query (Sec. 3.4.2). The full procedure is outlined in Algorithm 1.

#### 3.4.1 Stage I: Structured Bank Construction

The goal of stage I is to convert a set of raw context observations into a structured memory that can grow with the user. Rather than treating the $k$ context items in $\mathcal{V}_c$ as independent prompt examples, we view them as compounding evidence about the same underlying individual. The bank $\mathcal{B}$ is therefore represented as a set of entries $(\tau, d, e)$. Here, $\tau$ denotes the memory type, taking values in $\{\text{APPEARANCE}, \text{OWNED\_OBJECTS}, \text{BEHAVIOR}\}$ and broadly aligned with the three benchmark axes. $d$ is a natural-language memory descriptor grounded by the corresponding visual evidence $e$. Depending on the memory type, $e$ takes two forms: a single supporting frame for APPEARANCE and OWNED_OBJECTS to capture their static properties, or a video span for BEHAVIOR to preserve the temporal structure.

The bank is updated sequentially as new context items are processed. EXTRACT converts each item into a set of candidate memory entries $\mathcal{M}_i$; MERGE compares those candidates against the current bank and applies one of four updates. ADD creates a new entry, CONFIRM accumulates support for an existing entry, REVISE refines the descriptor when the new observation makes it more precise, and RETRACT removes an entry from the bank when later extracted cues suggest it is unreliable, such as a transient cue or an earlier visual misread. In this way, Stage I turns an expanding visual history into a compact, evidence-linked memory rather than a naive concatenation of prompt examples.

#### 3.4.2 Stage II: Query-Adaptive Evidence Selection

With the memory bank $\mathcal{B}$ established, the subsequent challenge is to find the right information for a given query. Because the bank aggregates rich appearance cues, personal objects, and behavioral routines from multiple clips, supplying a model with the entirety of this raw visual data is highly inefficient. At the same time, relying solely on the textual view of the bank is inherently lossy and sacrifices the precise visual details required for fine-grained personalization. To resolve this, we use each natural-language memory descriptor as a textual index into stored visual evidence. The model first surveys the text view $T_{\mathcal{B}}$ to identify a selected set of entry IDs $\mathcal{I}$ relevant to the current query. It then receives a hybrid view $H_{\mathcal{I}}$ that inlines visual evidence only for those selected entries. This process yields a query-tailored memory view.

We implement this as a two-step inference process. First, during an initial text triage, the LMM is presented with $T_{\mathcal{B}}$, listing $\tau$, $d$, and a stable ID for each entry alongside the query $\mathcal{L}_q$ and $\mathcal{V}_q$. Based on this lightweight summary, the model either answers the query directly or issues a tool call requesting the visual evidence $e$ for specific entry IDs. Second, we execute selective visual verification. If a tool call was issued, the bank is re-rendered as $H_{\mathcal{I}}$, which inlines the supporting frames or video spans strictly for the requested IDs while leaving the rest of the bank as text. The model then produces its final answer based on this hybrid context. By treating evidence selection as an explicit reasoning decision, our design ensures the model only loads visual evidence into the heavier context window when necessary.

Together, the two stages turn the diagnostic findings in Sec. 3.3 into a concrete Personal VCL baseline. Stage I tackles the scaling paradox, by shifting the unit of context from isolated clips to persistent, evidence-linked personal cues. Stage II addresses the modality paradox, by replacing passive visual concatenation with query-conditioned selection, which preserves the benefit of textual organization while retaining access to the supporting visual evidence. By establishing this structured approach, we aim to provide a principled starting point for subsequent research in Personal VCL.

## 4 Experiments

### 4.1 Setup

**Evaluation.** We evaluate the Agentic Context Bank as a strong inference-time baseline for Personal VCL across the full Personal-VCL-Bench suite using Gemma-4-31B [22], a representative strong open-weight LMM. Since EgoWearer identification integrates all three axes and places the strongest demand on context utilization, we further evaluate this task across Gemma-4-31B [22], Gemini-3-Flash [27], and GPT-5.4-mini [57] to test whether the baseline generalizes across LMM backbones. Implementation details are provided in Supp. C.

Table 2: Task-wise results of the Agentic Context Bank on Gemma-4-31B [22] across Personal-VCL-Bench (per-task accuracy, %). The comparisons span the main ways current LMMs are given context: No-context for query-only prompting, Language-ctx for textual descriptions of each context item [80, 74, 66], and Visual-ctx for direct prompting with the raw context images or clips [85, 38, 19]. Absolute gain is measured over the stronger of Language-ctx and Visual-ctx at $k_{\max}$.

| Context regime | Persons | | Objects | | Behavior | | EgoWearer |
|---|---|---|---|---|---|---|---|
| | ID | Rel | ID | Det | Err | QA | ID |
| No-context | 49.25 | 22.40 | 51.08 | 57.97 | 52.43 | 29.59 | 50.13 |
| Language-ctx ($k_{\max}$) | 78.00 | 49.60 | 78.75 | 71.74 | 58.11 | 34.69 | 52.99 |
| Visual-ctx ($k_{\max}$) | 75.75 | 46.40 | 75.34 | 65.22 | 59.46 | 27.55 | 55.36 |
| Ours | **83.25** | **51.20** | **82.52** | **73.19** | **66.22** | **35.71** | **61.60** |
| (Gain) | (+5.25) | (+1.60) | (+3.77) | (+1.45) | (+6.76) | (+1.02) | (+6.24) |

Table 3: EgoWearer Identification as an integrated Personal VCL test (accuracy, %). We evaluate the Agentic Context Bank across three LMM backbones and ablate stage-II evidence selection by comparing adaptive evidence selection against descriptors-only and all-evidence variants. The consistent gains indicate that Personal VCL requires structured and selective access to visual context.

| | Method | GPT-5.4-mini [57] | Gemma-4-31B [22] | Gemini-3-Flash [27] |
|---|---|---|---|---|
| Baselines | No-context | 49.83 | 50.13 | 49.47 |
| | Language-ctx ($k_{\max}$) | 49.86 | 52.99 | 55.26 |
| | Visual-ctx ($k_{\max}$) | 51.47 | 55.36 | 55.74 |
| Ours | Descriptors only | 50.69 | 51.97 | 55.28 |
| | All evidence | 51.21 | 57.29 | 67.86 |
| | Adaptive evidence | **53.77** | **61.60** | **72.82** |

**Baselines.** We compare against standard context prompting baselines. The language-context regime converts each item in $\mathcal{V}_c$ into a textual description and feeds only the descriptions to the LMM, a verbalize-and-answer pipeline widely used in long-form video understanding [80, 74, 66]. The visual-context regime concatenates the raw visual tokens of all items in $\mathcal{V}_c$ into the model prompt, the dominant setup in multi-image and many-shot visual in-context learning [85, 38, 19]. These comparisons place the Agentic Context Bank against the main existing choices for visual context utilization: no context, text summaries, or flat visual concatenation.

### 4.2 Results

**Task-wise results.** Tab. 2 evaluates the Agentic Context Bank across the full Personal-VCL-Bench suite. Across persons, objects, behavior, and EgoWearer identification, the bank improves over language-context and visual-context prompting baselines. The gains are largest on tasks that require aggregating or comparing personal evidence, such as EgoWearer identification. These results support our central claim that visual personalization requires not only access to relevant context, but also a better way to organize and use it.

**Cross-backbone analysis.** Tab. 3 studies EgoWearer identification, the benchmark capstone that requires integrating appearance, owned-objects, and behavioral cues. The full Agentic Context Bank consistently outperforms standard context prompting across all evaluated backbones, with the largest gain reaching 17.1% over the strongest standard baseline on Gemini-3-Flash [27]. These results indicate that the benefit of structured, query-adaptive memory is not tied to a single model family.

**Stage-II ablation.** The lower block of Tab. 3 ablates the design of query-adaptive evidence selection. We compare against two ways of querying the same bank. The first uses only the text view $T_{\mathcal{B}}$, exposing the structured memory descriptors without any stored visual evidence. The second exposes visual evidence for all bank entries, corresponding to an exhaustive version of the hybrid view. Both variants underperform adaptive selection, showing that the gain comes not only from building a better memory, but also from using it adaptively at query time. See Supp. D for stage-I ablation and additional analysis.

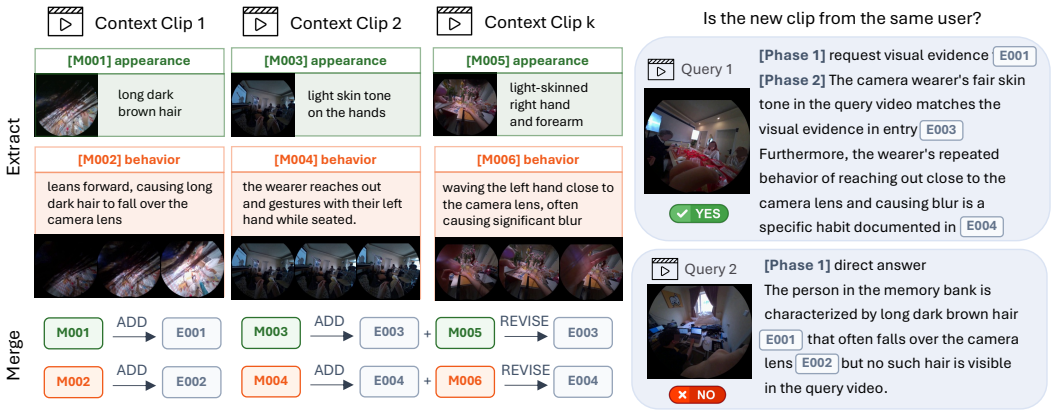


Figure 3: Qualitative example of the full Agentic Context Bank pipeline. The left side shows stage-I bank construction: from each context clip, the model extracts candidate memory entries (e.g., M001, M002) describing visually-grounded appearance or behavior cues. Through a merging step, these candidates are consolidated into stable evidence-linked bank entries (e.g., E001, E002), which preserve the supporting visual evidence for later inspection. The right side shows stage-II query-time use. Given a new query clip, the model first reasons over the text view of the bank and either requests selected evidence entries for visual verification or answers directly. In Query 1, the model requests relevant evidence and confirms the same wearer; in Query 2, the text descriptors already contradict the query clip, allowing a direct negative decision. We encourage readers to view the Supp. video for full temporal context.

**Qualitative examples.** Fig. 3 provides an example of the full bank lifecycle. From several context clips, the model extracts typed memory entries such as hand appearance and motion patterns. The merge process then updates the bank by adding new cues and revising overlapping ones, producing a compact memory rather than a flat list of clips. At inference time, the same bank supports different reasoning paths: the first query triggers targeted visual verification of relevant entries, while the second can be answered directly from the textual memory because the visible wearer contradicts stored identity cues.

**Limitations.** This work studies context utilization in isolation, assuming relevant moments are already retrieved from a continuous personal history. Because prior research has dedicated significant attention to the context acquisition problem, we enforce this specific boundary to demonstrate that frontier LMMs still fail even under this simplified condition (Sec. 3.3). Extending our framework to jointly evaluate retrieval and reasoning is an important next step. Furthermore, our Agentic Context Bank is intended as a diagnostic baseline for this context utilization problem; its improvements identify useful ingredients, while its remaining errors (see Supp. D for failure cases) highlight a clear need for continued innovation in this domain.

## 5 Conclusion

The path from generic LMMs to genuinely personal assistants runs through a capability that pre-training cannot supply by construction: reasoning from a user's own visual history, over knowledge that is unique to that user. We introduce Personal VCL to formalize this capability, alongside Personal-VCL-Bench to systematically evaluate LMM performance. Our investigation diagnoses a severe context utilization gap in frontier LMMs, characterized by a modality paradox (an over-reliance on the lossy text modality) and a scaling paradox (an inability to reliably aggregate expanding visual context). To address both, we propose the Agentic Context Bank. This inference-time framework aggregates context items into a structured, self-refining memory that is consulted via query-adaptive evidence selection, greatly narrowing the performance gap.

Personal VCL is, we believe, one of the defining capabilities on the trajectory towards personalized AI systems. We hope the problem formulation, benchmark, diagnostic findings, and baseline method this paper offers are a useful starting point for that effort. To support the broader research community, we will publicly release all aspects of Personal-VCL-Bench and our baseline implementations.

## References


[1] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. MyVLM: Personalizing VLMs for user-specific queries. *arXiv preprint arXiv:2403.14599*, 2024.

[2] Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, and Jiwen Cao. MC-LLaVA: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

[3] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, and Gaole Dai. UniCTokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025.

[4] Ruichuan An, Kai Zeng, Ming Lu, Sihan Yang, Renrui Zhang, Huitong Ji, Hao Liang, and Wentao Zhang. Concept-as-tree: A controllable synthetic data framework makes stronger personalized VLMs. *arXiv preprint arXiv:2503.12999*, 2025.

[5] Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. *arXiv preprint arXiv:2507.02001*, 2025.

[6] Huiyu Bai, Runze Wang, Zhuoyun Du, Yiyang Zhao, Fengji Zhang, Haoyu Chen, Xiaoyong Zhu, Bo Zheng, and Xuejiao Zhao. Online-PVLM: Advancing personalized VLMs with online concept learning. *arXiv preprint arXiv:2511.20056*, 2025.

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[8] Jing Bi, Yunlong Huang, Lianggong Wang, and Jiebo Luo. EAGLE: Egocentric AGgregated language-video engine. *arXiv preprint arXiv:2409.17523*, 2024.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[10] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[11] Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgöz, Shreyas Hampali, Eric Sauser, Shugao Ma, Angela Yao, and Fadime Sener. Memory-efficient streaming VideoLLMs for real-time procedural video understanding. *arXiv preprint arXiv:2504.13915*, 2025.

[12] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision (ECCV)*, 2022.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)*, 2022.

[14] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[15] Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. ReWind: Understanding long videos with instructed learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2023.

[17] Yuhao Dong, Shulin Tian, Shuai Liu, Shuangrui Ding, Yuhang Zang, and Xiao Dong. Demo-ICL: In-context learning for procedural video knowledge acquisition. *arXiv preprint arXiv:2602.08439*, 2026.

[18] Shihan Dou, Ming Zhang, Zhangyue Yin, et al. CL-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

[19] Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. In *European Conference on Computer Vision (ECCV)*, 2024.

[20] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Andreas Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*, 2023.

[21] Chenyou Fan. EgoVQA: An egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[22] Clement Farabet and Olivier Lacombe. Gemma 4: Byte for byte, the most capable open models. Technical report, Google, April 2026.

[23] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[24] Ryo Fujii, Hideo Saito, and Ryo Hachiuma. VIOLA: Towards video in-context learning with minimal annotations. *arXiv preprint arXiv:2601.15549*, 2026.

[25] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[26] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[27] Gemini Team, Google DeepMind. Gemini 3 Flash model card. Technical report, Google DeepMind, December 2025.

[28] Gemini Team, Google DeepMind. Gemini 3.1 Pro model card. Technical report, Google DeepMind, February 2026.

[29] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[31] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. RAP: Retrieval-augmented personalization for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[32] Rongpei Hong, Jian Lang, Ting Zhong, Yong Wang, and Fan Zhou. TAMEing long contexts in personalization: Towards training-free and state-aware MLLM personalized assistant. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. arXiv:2512.21616.

[33] Yedid Hoshen and Shmuel Peleg. An egocentric look at video photographer identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4284–4292, 2016.

[34] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong Lu, Yali Wang, Limin Wang, and Yu Qiao. EgoExoLearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Mingfang Zhang, Lijin Yang, Zheng Nie, Jinyao Liu, Guoshun Fan, Dechen Lin, Fang Fang, Kunpeng Li, Chang Yuan, Xinyuan Chen, Yaohui Wang, Yali Wang, Yu Qiao, and Limin Wang. An egocentric vision-language model based portable real-time smart assistant. *arXiv preprint arXiv:2503.04250*, 2025.

[36] Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. BIMBA: Selective-scan compression for long-range video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[37] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems*, 2022.

[38] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, and Jonathan Chen. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024.

[39] Jaeik Kim, Woojin Kim, Woohyeon Park, and Jaeyoung Do. MMPB: It's time for multi-modal personalization. *arXiv preprint arXiv:2509.22820*, 2025.

[40] Kangsan Kim, Geon Park, Youngwan Lee, and Sung Ju Hwang. MA-EgoQA: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827*, 2026.

[41] Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. VideoICL: Confidence-based iterative in-context learning for out-of-distribution video understanding. *arXiv preprint arXiv:2412.02186*, 2024.

[42] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

[43] Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, 2021.

[44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

[45] Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. LensWalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026.

[46] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[47] Yijiang Li et al. EgoPrivacy: What your first-person camera says about you? In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.12258.

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[49] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria Everyday Activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

[50] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, 2023.

[51] Zaira Manigrasso, Matteo Milani, and Rita Cucchiara. Online episodic memory visual query localization with egocentric streaming object memory. *arXiv preprint arXiv:2411.16934*, 2024.

[52] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[53] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jake Hofman, and Jennifer Neville. PEARL: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*, 2024.

[54] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your personalized language and vision assistant. In *Advances in Neural Information Processing Systems*, 2024.

[55] Yeongtak Oh, Dohyun Chung, Juhyeon Shin, Sangha Park, Johan Barthelemy, Jisoo Mok, and Sungroh Yoon. RePIC: Reinforced post-training for personalizing multi-modal language models. *arXiv preprint arXiv:2506.18369*, 2025.

[56] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[57] OpenAI. GPT-5.4 Thinking system card. Technical report, OpenAI, March 2026.

[58] Rohith Peddi, Shivvrat Arya Tirumala, Mohammad Khan, Yufei Ji, Tushar Sridhar, Vibhav Gogate Sridhar, and Nicholas Ruozzi. CaptainCook4D: A dataset for understanding errors in procedural activities. *arXiv preprint arXiv:2312.14556*, 2023.

[59] Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*, 2024.

[60] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*, 2024.

[61] Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, and Chai. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*, 2026.

[62] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2024.

[63] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

[64] Alireza Salemi and Hamed Zamani. LaMP-QA: A benchmark for personalized long-form question answering. *arXiv preprint arXiv:2506.00137*, 2025.

[65] Soroush Seifi, Vaggelis Dorovatas, Matteo Cassinelli, Fabien Despinoy, Daniel Olmeda Reino, and Rahaf Aljundi. Personalization toolkit: Training free personalization of large vision language models. *arXiv preprint arXiv:2502.02452*, 2025.

[66] Junxiao Shen, John Dudley, and Per Ola Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. *arXiv preprint arXiv:2308.05822*, 2023.

[67] Yufei Shi, Weilong Yan, Gang Xu, Yumeng Li, Yucheng Chen, Zhenxi Li, Fei Richard Yu, Ming Li, and Si Yong Yeo. PVChat: Personalized video chat with one-shot learning. *arXiv preprint arXiv:2503.17069*, 2025.

[68] Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[69] Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*, 2024.

[70] Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, and Meng Jiang. Instant personalized large language model adaptation via hypernetwork. *arXiv preprint arXiv:2510.16282*, 2025.

[71] Daksh Thapar, Aditya Nigam, and Chetan Arora. Is sharing of egocentric video giving away your biometric signature? In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[72] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[73] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, 2024.

[74] Ying Wang, Yanlai He, Cuiling Wang, Huaiyu Bian, and Jianlong Chen. LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023.

[75] Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025.

[76] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, 2024.

[77] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. *arXiv preprint arXiv:2401.00789*, 2024.

[78] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, et al. EgoLife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.

[79] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting in-context learning in vision-language models for videos through curated data distributional properties. *arXiv preprint arXiv:2311.17041*, 2023.

[80] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[81] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *AAAI Conference on Artificial Intelligence*, 2020.

[82] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024.

[83] Yuanhong Zheng, Ruichuan An, Xiaopeng Lin, Yuxing Liu, Sihan Yang, Huanyu Zhang, Haodong Li, Qintong Zhang, Renrui Zhang, Guopeng Li, Yifan Zhang, Yuheng Li, and Wentao Zhang. PEARL: Personalized streaming video understanding model. *arXiv preprint arXiv:2603.20422*, 2026.

[84] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: A comprehensive benchmark for multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[85] Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VL-ICL bench: The devil in the details of multimodal in-context learning. *arXiv preprint arXiv:2403.13164*, 2024.

# Supplementary Material

## Contents



## A Personal-VCL-Bench

### A.1 Task summary

Personal-VCL-Bench is built by repurposing source videos from EgoLife [78], Ego4D [29], and CaptainCook4D [58] into personalized context-query tasks. The source datasets are used under their original terms: EgoLife is listed as MIT-licensed in its Hugging Face release, Ego4D is governed by the Ego4D License Agreement, and CaptainCook4D is released under Apache License 2.0.

The benchmark is constructed to isolate a central question: given curated visual evidence from a user's history, can an LMM use that evidence to answer a new user-specific visual query? Towards this end, we design seven tasks spanning the major forms of personal visual knowledge an assistant must acquire: familiar people, personally relevant objects, individualized behavior, and the wearer's own identity. Tab. 4 reports the task organization, context-query modality, task size, and the average number of context images or clips available for each query. Our design stresses two core challenges of Personal VCL: fine-grained visual understanding, and the need to aggregate evidence from multiple context observations. Together, these properties make the benchmark a targeted testbed for diagnosing Personal VCL in LMMs.

### A.2 Task construction

**Persons.** The persons axis is designed to test whether Personal VCL can support social understanding, beginning with visual identity binding and extending to relationship reasoning. EgoLife [78] provides multi-day egocentric recordings from a shared household. We focus on five housemates observed from the camera wearer (Jake)'s perspective: Alice, Katrina, Lucia, Shure, and Tasha. For each housemate, we manually select clear Day-1 reference images to form a high-quality visual context, ensuring that the context reliably specifies the target identity. The single-identity recognition task then asks whether the named housemate appears in a query image from Day 6 or 7. This temporal gap makes the task a test of persistent identity recognition rather than near-duplicate matching, yielding 250 queries. We further evaluate whether such identity context can support higher-level social

Table 4: Personal-VCL-Bench task summary. Seven tasks span three axes of personal visual knowledge (persons, objects, behavior) plus the EgoWearer Identification capstone. $\overline{|\mathcal{V}_c|}$ reports the average number of context images or clips per query. 🖼 = image, 🎞 = video.

| Axis | Task ID | Task | Source | $\mathcal{V}_c \to \mathcal{V}_q$ | $\overline{\lvert\mathcal{V}_c\rvert}$ | # |
|---|---|---|---|---|---|---|
| Persons | PerID | Single-identity recognition | EgoLife [78] | 🖼 → 🖼 | 5.00 | 250 |
| | PerRel | Relationship reasoning | EgoLife [78] | 🖼 → 🎞 | 25.00 | 125 |
| Objects | ObjID | Object identity recognition | Ego4D [29] | 🖼 → 🎞 | 1.42 | 660 |
| | ObjDet | Personalized object detection | Ego4D [29] | 🖼 → 🖼 | 1.72 | 138 |
| Behavior | BehErr | Procedural error detection | CaptainCook4D [58] | 🎞 → 🎞 | 1.98 | 370 |
| | BehQA | Procedural question answering | CaptainCook4D [58] | 🎞 → 🎞 | 2.40 | 98 |
| EgoWearer | EgoID | EgoWearer Identification | EgoLife [78] | 🎞 → 🎞 | 5.00 | 614 |

reasoning using the RelationQA subset from EgoLifeQA [78]. RelationQA already defines four-way questions over Jake's egocentric video, but does not provide the curated identity context needed to isolate visual context utilization. We therefore pair each question with our manually selected reference galleries for the household members, requiring the model to resolve participant identities from personal visual context before answering the social question. This task contains 125 queries.

**Objects.** The objects axis evaluates whether a model can ground a personally specified object instance, rather than merely recognize an object category. We build this axis from Ego4D Visual Queries [29] by using its object tracks and bounding-box annotations as raw visual evidence, then redesigning them into Personal VCL context-query tasks. Across eight everyday object categories, bottle, bowl, box, bucket, chair, container, cup, and pliers, we form reference galleries for individual object instances. In instance recognition, the model is given one to five reference images of a target object, such as "my bottle," and must decide whether the same instance appears in a query video. Negative queries are drawn from other instances of the same category whenever possible, forcing instance-level discrimination rather than category recognition. This task contains 660 queries over 89 object instances. In spatial detection, we use the same personalized reference setup but ask the model to localize the target instance in a query frame, using Ego4D's bounding-box annotations for evaluation. This yields 138 positive localization queries over 88 object instances. Together, the two tasks turn object-level egocentric annotations into a test of personalized object grounding: recognizing and localizing the specific object tied to the user.

**Behavior.** The behavior axis targets a form of personal knowledge that is difficult to capture with static concepts: how a user performs a routine action. We redesign CaptainCook4D [58] into this Personal VCL setting by constructing same-user, same-step context-query pairs. For procedural error detection, the context clips show the participant's correct execution of a step, while the query clip shows the same participant performing that step again. The model must decide whether the query deviates from the demonstrated baseline. Because the context and query involve the same user and the same procedural step, the task focuses on personalized deviation detection rather than generic mistake recognition. This task contains 370 balanced binary queries. For procedural question answering, a human annotator compares the same context and query videos and identifies fine-grained attributes that remain consistent or differ in the user's execution. We formulate these annotations into 98 four-way multiple-choice questions (MCQ), challenging the model to tell how an execution changes relative to personal visual context.

**EgoWearer identification.** This final task asks whether an LMM can recognize the continuity of the camera wearer from first-person visual experience. EgoLife [78] provides a particularly strong basis for this task because it records six subjects over multiple days in the same shared apartment, producing natural overlap in scenes, activities, and household objects. Using the dataset's action annotations, we design each example around five reference clips of one wearer performing a particular action. The query clip shows either the same wearer or another individual performing the same action, and the model must decide whether the query wearer matches the reference wearer. Because negatives are drawn from another person doing the same action in an overlapping environment, scene and action cues are intentionally controlled; the model must instead rely on fine-grained identity evidence such as hands, clothing, personal objects, and motion patterns. We build 614 queries spanning 15 daily actions, with 96 manually curated behavior-centric queries whose context clips exhibit clean and stable person-specific motion patterns. The resulting task provides a demanding testbed for multi-context aggregation and fine-grained egocentric identity reasoning.

### A.3 Evaluation

We choose metrics according to the output format of each task. For binary recognition tasks, PerID, ObjID, BehErr, and EgoID, we report macro accuracy, averaging Yes and No accuracy so that class imbalance does not reward majority-class predictions. For four-way reasoning tasks, PerRel and BehQA, we report standard multiple-choice accuracy, counting invalid or unparseable outputs as incorrect. For personalized object localization, ObjDet, we evaluate the predicted bounding box by intersection-over-union and report Acc@IoU $\geq 0.5$. For EgoID, we report a single score over the full 614-query evaluation set: class accuracies are computed after pooling the general and behavior-centric subsets, then macro-averaged.

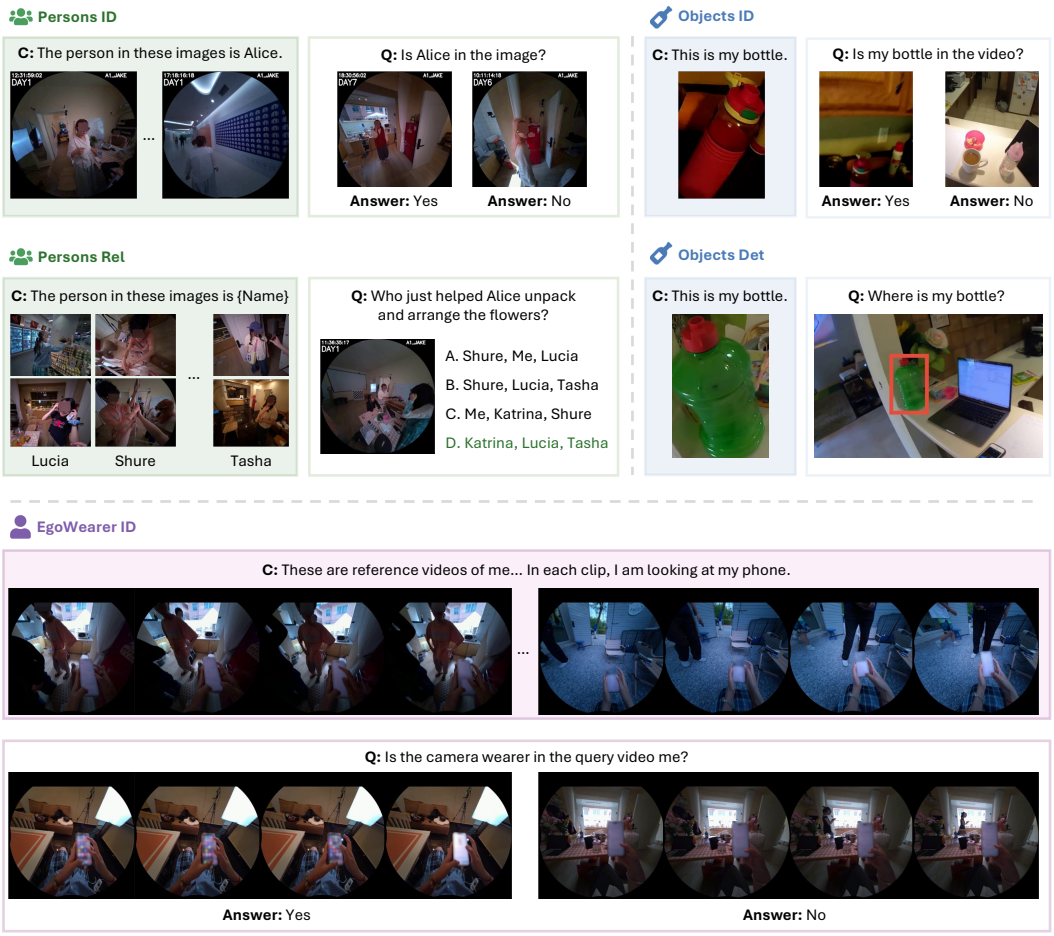


Figure 4: Examples from the Persons, Objects, and EgoWearer identification tasks in Personal-VCL-Bench, built from EgoLife [78] and Ego4D [29]. C denotes context, and Q denotes the query.

## B Extended Diagnostic Results

We provide the full diagnostic table in Tab. 5, extending the representative results in the main paper to all seven LMMs. The evaluation follows the same five-regime protocol for every model: no-context, language-context with one item, language-context with all items, visual-context with one item, and visual-context with all items.

The results support three conclusions. First, Personal-VCL-Bench largely resists generic pretrained knowledge. Identity-style tasks such as PerID, ObjID, and EgoID remain close to chance for most models, while behavior and relation tasks can contain partial signal in the query itself. This makes the subsequent context comparisons more important: they measure whether language or visual context helps models move beyond generic query understanding toward genuine personal visual context utilization. Second, the modality comparison exposes a gap in native visual context utilization. Visual context can be powerful, particularly for person and object identity, but it is not reliably dominant; language descriptions often match or outperform raw visual evidence despite discarding visual detail. Third, the $k{=}1$ versus $k_{\max}$ comparison shows that current LMMs do not reliably aggregate personal context. More reference images or clips can help when the task resembles visual matching, but the benefit becomes unstable for behavior, localization, and EgoWearer identification. This pattern is central to our motivation: Personal VCL requires not only retrieving relevant visual history, but also organizing and selecting from it in a way that current prompt concatenation does not provide.

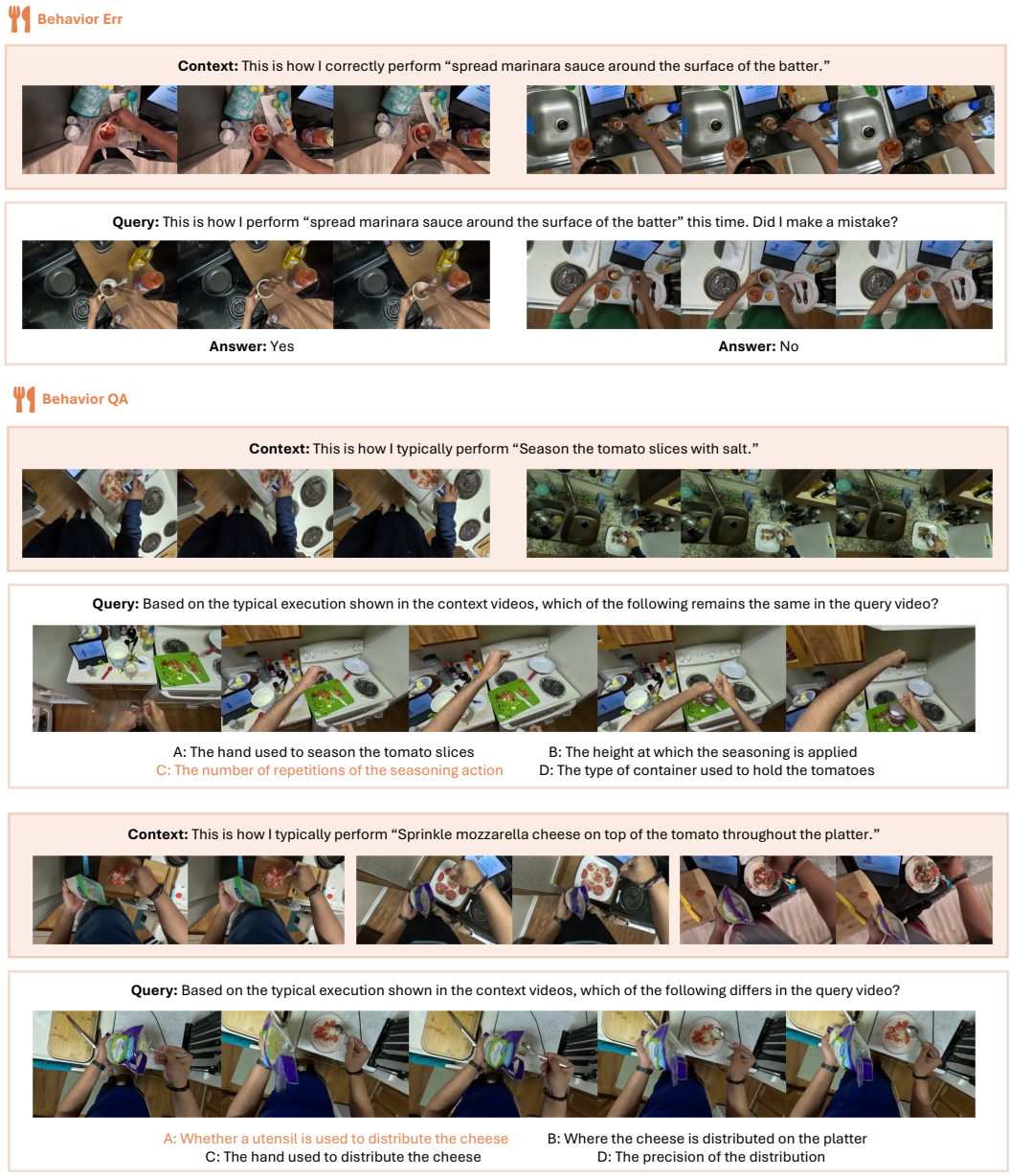


Figure 5: Examples from the Behavior tasks in Personal-VCL-Bench, built from CaptainCook4D [58].

Table 5: Full diagnostic results on Personal-VCL-Bench across five context regimes for all seven evaluated LMMs (per-task accuracy, %); extends Tab. 1. **Bold** = best within a model's five rows; underline = best across all models. A random baseline is not directly applicable to the continuous bounding box localization required for ObjDet (Acc@IoU≥0.5) and is marked as '–'. In visual-context rows, cell shading indicates whether visual context improves, hurts, or ties against the matched language-context row. In Visual-ctx ($k_{\max}$), superscript markers indicate whether increasing visual context from $k$=1 to $k_{\max}$ improves, drops, or ties.

| Context regime | Persons | | Objects | | Behavior | | EgoWearer |
|---|---|---|---|---|---|---|---|
| | ID | Rel | ID | Det | Err | QA | ID |
| Random guess | 50.00 | 25.00 | 50.00 | – | 50.00 | 25.00 | 50.00 |
| Qwen3-VL-8B [7] | | | | | | | |
| No-context | 50.50 | 36.00 | 49.06 | 68.12 | 49.46 | 19.39 | 49.83 |
| Language-ctx ($k$=1) | 51.75 | 29.60 | 65.59 | 76.09 | **57.57** | **24.49** | 52.57 |
| Language-ctx ($k_{\max}$) | 62.75 | **48.80** | 67.25 | **77.54** | 54.86 | 22.45 | **54.53** |
| Visual-ctx ($k$=1) | 62.75 | 46.40 | 70.99 | **77.54** | **57.57** | 21.43 | 50.00 |
| Visual-ctx ($k_{\max}$) | **85.25**↑ | 46.40~ | **73.67**↑ | 76.81~ | 55.14↓ | 22.45↑ | 50.00~ |
| Qwen3-VL-32B [7] | | | | | | | |
| No-context | 49.25 | 32.80 | 48.47 | 61.59 | 48.38 | **34.69** | 49.65 |
| Language-ctx ($k$=1) | 50.00 | 34.40 | 70.29 | **78.26** | **51.08** | 29.59 | 52.49 |
| Language-ctx ($k_{\max}$) | 70.25 | 49.60 | 71.02 | 76.09 | 49.73 | 26.53 | 55.34 |
| Visual-ctx ($k$=1) | 67.50 | 43.20 | 78.67 | 39.86 | 50.54 | 26.53 | 53.02 |
| Visual-ctx ($k_{\max}$) | **85.75**↑ | **56.00**↑ | **79.99**↑ | 33.33↓ | 48.92↓ | 24.49↓ | **62.52**↑ |
| InternVL-3.5-8B [72] | | | | | | | |
| No-context | 54.25 | 33.60 | 49.34 | **19.57** | 51.08 | 11.22 | 47.83 |
| Language-ctx ($k$=1) | 54.50 | 36.00 | 62.86 | 13.77 | 53.24 | **19.39** | 53.36 |
| Language-ctx ($k_{\max}$) | 59.50 | 36.00 | **63.05** | 14.49 | 52.97 | 17.35 | **53.92** |
| Visual-ctx ($k$=1) | 54.75 | **39.20** | 50.29 | 10.87 | **54.86** | 12.24 | 50.25 |
| Visual-ctx ($k_{\max}$) | **62.50**↑ | 38.40~ | 50.29~ | 7.25↓ | 54.05~ | 8.16↓ | 49.83~ |
| InternVL-3.5-38B [72] | | | | | | | |
| No-context | 53.75 | 38.40 | 49.95 | **13.77** | 50.00 | 22.45 | 50.00 |
| Language-ctx ($k$=1) | 54.50 | 37.60 | 65.41 | 10.87 | **52.97** | 28.57 | 52.80 |
| Language-ctx ($k_{\max}$) | 75.75 | **50.40** | **65.76** | 10.14 | 51.35 | **31.63** | **53.78** |
| Visual-ctx ($k$=1) | 77.25 | 42.40 | 52.02 | 12.32 | 50.54 | 17.35 | 49.84 |
| Visual-ctx ($k_{\max}$) | **80.50**↑ | 47.20↑ | 51.79~ | 9.42↓ | 50.00~ | 19.39↑ | 50.17~ |
| Gemma-4-31B [22] | | | | | | | |
| No-context | 49.25 | 22.40 | 51.08 | 57.97 | 52.43 | 29.59 | 50.13 |
| Language-ctx ($k$=1) | 57.50 | 20.80 | 78.33 | **72.46** | 54.86 | 33.67 | 54.52 |
| Language-ctx ($k_{\max}$) | **78.00** | **49.60** | **78.75** | 71.74 | 58.11 | **34.69** | 52.99 |
| Visual-ctx ($k$=1) | 65.50 | 30.40 | 73.26 | 63.77 | **59.46** | 32.65 | **56.66** |
| Visual-ctx ($k_{\max}$) | 75.75↑ | 46.40↑ | 75.34↑ | 65.22↑ | **59.46**~ | 27.55↓ | 55.36↓ |
| Gemini-3-Flash [27] | | | | | | | |
| No-context | 55.75 | 42.40 | 49.32 | 64.49 | 57.84 | 37.76 | 49.47 |
| Language-ctx ($k$=1) | 69.25 | 37.60 | 80.55 | 79.71 | 55.41 | 45.92 | 53.99 |
| Language-ctx ($k_{\max}$) | 81.75 | **59.20** | 82.35 | 78.99 | 54.59 | 48.98 | 55.26 |
| Visual-ctx ($k$=1) | 67.50 | 44.00 | 82.21 | **85.51** | **68.92** | 47.96 | 51.99 |
| Visual-ctx ($k_{\max}$) | **88.00**↑ | 54.40↑ | **82.47**~ | 82.61↓ | 67.03↓ | **61.22**↑ | **55.74**↑ |
| Gemini-3.1-Pro [28] | | | | | | | |
| No-context | 54.50 | 43.20 | 51.64 | 72.46 | 66.22 | 35.71 | 49.35 |
| Language-ctx ($k$=1) | 52.75 | 44.00 | 67.54 | 82.61 | 58.38 | 47.96 | 48.73 |
| Language-ctx ($k_{\max}$) | 73.50 | 45.60 | 78.28 | **83.33** | 56.22 | 51.02 | 49.75 |
| Visual-ctx ($k$=1) | 83.75 | 50.40 | **83.63** | 55.07 | 64.32 | 61.22 | 55.90 |
| Visual-ctx ($k_{\max}$) | **88.75**↑ | **60.00**↑ | 82.14↓ | 50.00↓ | **68.11**↑ | **65.31**↑ | **58.16**↑ |

# C Implementation Details

## C.1 Context prompting baselines

We evaluate three direct prompting baselines that vary only in how personal context is represented. The no-context baseline receives only the query media and the task question. The visual-context baseline presents the selected reference images or clips before the query. The language-context baseline first asks the same model to describe each reference item, then presents the task using those descriptions in place of the raw reference media. For both context baselines, we report a single-reference condition and a full-reference condition, denoted $k{=}1$ and $k_{\max}$. Across these settings, the task question is unchanged; the comparison isolates whether the model can use personal context, and whether that context is more effective as raw visual evidence or as text. We uniformly sample 16 frames for all video inputs, except for Gemini models, which process native video via their API. These identical settings are applied when generating language-context descriptions.

**Visual context baseline prompts.** The task-specific visual-context prompts are shown below.

**Persons**

For single-identity recognition, the context prompt is "The person in these images is `[person]`" followed by the query "Is `[person]` in this image?" For relationship reasoning, each household member is introduced with the same identity prompt, and the query is the RelationQA question.

**Objects**

For instance recognition and spatial detection, the shared context prompt is "This is my `[object]`." The recognition query asks "Is my `[object]` in the video?" The detection query asks "Where is my `[object]`? Return the bounding box coordinates."

**Behavior**

For procedural error detection, the context prompt is "This is how I correctly perform `[step]`." The query asks "This is how I perform `[step]` this time. Did I make a mistake?" For procedural question answering, the context prompt is "This is how I typically perform `[step]`." The query is the human-written multiple-choice question about which aspect differs from, or remains consistent with, the context execution, followed by four answer choices.

**EgoWearer**

For EgoWearer identification, the context prompt is "These are reference videos of me, recorded with a head-mounted fisheye camera in a shared apartment with my housemates. In each clip I am `[action]`." The query asks: "All reference and query clips share the same camera rig, the same apartment, the same housemates, and the same action, so none of those are evidence of identity. Look for subtler, person-specific details that remain. Is the camera wearer in the query video me?"

**Language context baseline prompts.** The description prompts are shown below.

**Persons**

For persons, the description prompt is "These are images of `[person]`. Describe `[person]`'s physical appearance in one paragraph, focusing only on features that would help identify them: hair color and style, clothing, accessories such as glasses, jewelry, hats, or bags, body type, and any other distinguishing visual features. Do not describe the background, setting, or activities. Starting with '`[person]` is'." The generated descriptions replace the reference images as the identity context for both single-identity recognition and relationship reasoning.

**Objects**

For objects, the description prompt is "This is my `[object]`. Describe my `[object]` in one paragraph. Starting with 'It is'." The generated descriptions replace the reference images as the object context, framed as "This is my `[object]`. `[description]`."

**Behavior**

For behavior, the description prompt is "Describe how I perform `[step]`." The generated descriptions replace the reference clips as the procedural context, framed as "This is how I correctly perform `[step]`: `[description]`."

**EgoWearer**

For EgoWearer identification, each reference clip is described with "This is a first-person video clip of me `[action]`. Describe my appearance and my movements in detail." The generated descriptions replace the

reference clips as the wearer context, using the same apartment-and-action lead-in as the visual-context baseline, followed by "Below is a description of one such clip:" or "Below are descriptions of `[n]` such clips:" and the per-clip descriptions.

### C.2 Agentic Context Bank

**Stage I: bank construction.** In the multi-clip EgoWearer identification setting, each context clip is sampled into 16 uniformly spaced frames before the LMM extracts reusable personal cues about the camera wearer. Static cues, including appearance and owned objects, are grounded to a single supporting frame; motion cues are grounded to a temporal span. The extraction prompt is shown below.

**Extraction.**

"You are observing an egocentric (first-person) video clip from a head-mounted camera. The video has been sampled into `[N]` frames, labeled in temporal order. Extract distinctive cues about the camera wearer—observations that capture who this individual is and could be reused to reason about them in other contexts."

Static cues: what the wearer's body, clothing, wearables, or owned objects look like. Anchor each cue to one best frame.

Motion cues: how the wearer moves and acts. Anchor each cue to one temporal span.

Rules: one entry per distinct cue; do not bundle unrelated attributes; emit only cues grounded in a specific frame or span; prefer specific over vague descriptions; avoid restating the action label.

After extraction, candidates are reconciled with the current bank separately for each memory type. Each stored entry maintains a stable identifier, a descriptor, visual evidence, and support counts. Revisions are accepted only after a separate visual verification prompt. The merge and revision-verification prompts are shown below.

**Merge.**

"You are reconciling new `[category]` cues against an existing memory of a person."

Existing entries: `[active bank entries in this category]`.

New candidates: `[c_001]`, `[c_002]`, . . ., each with its supporting frame or temporal span.

Output: for each useful candidate, choose `ADD`, `CONFIRM`, `REVISE`, or `RETRACT`; silently drop redundant noise. Use `REVISE` only when the candidate refines the same underlying attribute.

**Revision verification.**

"You proposed the following REVISE operations. For each one, I am showing the visual evidence from both the existing entry and the new candidate. Verify that they refer to the same `[category]` attribute of the same person. If they do, confirm the REVISE; otherwise, withdraw it."

**Stage II: query-adaptive evidence.** At query time in this same wearer-recognition setting, the active bank is rendered as text grouped by appearance, owned objects, and behavior. The first call receives this text view together with 16 uniformly sampled query frames. If visual grounding is requested, the second call attaches evidence only for the requested entries: a supporting frame for static entries and sampled span frames for behavior entries. Across the reported agentic experiments, each requested behavior span is represented by up to four uniformly spaced frames for Gemini/GPT and by its two endpoints for Gemma. The two query-time prompts are shown below.

**Call 1: text triage.**

"These are reference videos of me, recorded with a head-mounted fisheye camera in a shared apartment with my housemates. In each clip I am `[action]`. Below is a structured memory of me, built from those clips:

`[bank text]`

Now here are frames from the query clip."

Question: "All reference and query clips share the same camera rig, the same apartment, the same housemates, and the same action, so none of those are evidence of identity. Look for subtler, person-specific details that remain. Is the camera wearer in the query video me?"

Decision: If you can decide from the text claims alone, answer Yes or No; otherwise, request the bank entries whose visual evidence would help you decide.

**Call 2: selective visual verification.**

"Here is the same structured memory of me (reference videos where I am `[action]`), now with visual evidence attached for the entries you requested:

`[bank text, with frames inlined only for requested entries]`

Visual evidence has been attached for the entries you requested: `[entry ids]`. Compare the attached visual evidence to the query video. Is the camera wearer in the query video me? Answer Yes or No and report the decisive entries."

**Task-wise extensions.** The preceding prompts describe the full two-stage Agentic Context Bank as instantiated for EgoWearer identification. For the task-wise results in Tab. 2, we adapt the bank to the structure of each benchmark axis. For entity-centric tasks in the persons and objects axes (PerID, PerRel, ObjID, ObjDet), there is little temporal structure to merge: each context item is already a reference observation of a known person or object. We therefore use a per-item description bank and focus on stage-II query-adaptive evidence selection. The model first triages the text bank, then selectively requests the corresponding visual evidence when the descriptions alone are insufficient. For the behavior axis, the relevant context is procedural rather than instance-level. We construct a phase-structured behavior bank from the reference executions and focus on evaluating the resulting stage-I bank. The task-wise prompt templates are shown below.

**Compute resources.** All reported experiments use frozen models at inference time. For local LMMs, we run inference on an NVIDIA H200 GPU cluster, with each inference job allocated one H200 GPU node. Proprietary models, including the Gemini and OpenAI families, are accessed via their official APIs.

**Persons** **Objects**

Call 1: "Here are text descriptions of `[reference images/clips]`:

`[e_001] [description 1]`
`[e_002] [description 2]`
...

Now here is the query: `[task question]`. Answer directly from the text descriptions, or request the specific entries whose visual evidence you need."

Call 2: "Here are the reference images or clips you requested. `[requested visual evidence]` Now, with this visual evidence alongside the query, answer the question: `[task question]`."

**Behavior**

BehErr: "Here is a record of how I perform `[step]`, broken into temporal phases (in order):

`[e_001] [phase 1]`
`[e_002] [phase 2]`
...

Now here are frames from a new clip where I perform the same action. Did I make a mistake? Answer Yes or No."

BehQA: "Here is a record of how I perform `[step]`, broken into temporal phases (in order):

`[e_001] [phase 1]`
`[e_002] [phase 2]`
...

Detailed description of how I typically perform this action across the reference clips: `[reference description]`

`[MCQ question]`"

## D Additional Results and Analysis

**Stage-I ablation.** Tab. 6 separates the effect of stage-I memory construction from the stage-II evidence-selection mechanism analyzed in the main paper. We consider three variants. First, we replace the structured bank with a flat visual-text memory, where each reference clip is represented by an independent description without extract-and-merge. Second, we perturb the order in which reference clips are processed during bank construction, testing whether the sequential update process is sensitive to context order. Third, we substitute the Gemma-built bank with a Gemini-built bank while still using Gemma for the final query-time decision. When the memory is reduced to flat per-clip descriptions, the same query-adaptive procedure obtains 51.11%, indicating that selective access

Table 6: Effect of stage-I bank construction on EgoWearer Identification. With Gemma-4-31B [22] as the query model, we compare flat visual-text memory, the full structured bank, an order-perturbed bank, and a bank constructed by a stronger builder. The full bank improves over flat memory, is stable to reference order, and benefits from stronger construction.

| Memory construction | Builder | EgoWearer |
|---|---|---|
| Flat visual-text memory | Gemma-4-31B | 51.11 |
| Full bank | Gemma-4-31B | 61.60 |
| Full bank (permuted order) | Gemma-4-31B | 62.11 |
| Full bank | Gemini-3-Flash | 65.53 |

alone is not sufficient if the underlying memory remains unstructured. The perturbed-order bank performs similarly at 62.11%, suggesting that the gain is not an ordering artifact. The Gemini-built bank further improves to 65.53%, indicating that stronger stage-I memory construction can directly benefit downstream Personal VCL.

Table 7: Stage-I bank-construction statistics on EgoWearer Identification. Cues/clip counts extracted candidate cues. Entries reports the mean final bank size, with A/O/B denoting appearance, owned-object, and behavior entries. Compression is the ratio between final bank entries and extracted candidates. Revision ops are the share of merge decisions that update an existing entry, and updated entries are final entries confirmed or revised after creation.

| Builder | Cues/clip | Entries | A/O/B | Comp. | Rev. ops | Updated |
|---|---|---|---|---|---|---|
| Gemini-3-Flash | 4.44 | 13.8 | 6.0/3.1/4.7 | 0.62 | 35.9% | 24.1% |
| Gemma-4-31B | 2.95 | 8.2 | 3.8/1.5/3.0 | 0.56 | 25.2% | 18.3% |
| GPT-5.4-mini | 6.92 | 18.4 | 6.7/5.7/5.9 | 0.53 | 38.1% | 26.5% |

Table 8: Stage-II query-time evidence statistics on EgoWearer Identification. Visual requests are the fraction of queries for which the model asks to inspect stored visual evidence. Requested and decisive entries are averaged over visual-evidence queries. A/O/B reports the percentage distribution over appearance, owned-object, and behavior entries.

| Builder | Vis. req. | Req. | Req. A/O/B | Dec. | Dec. A/O/B |
|---|---|---|---|---|---|
| Gemini-3-Flash | 95.8% | 3.63 | 81.8/15.1/3.0 | 2.67 | 79.2/15.2/5.5 |
| Gemma-4-31B | 96.1% | 2.86 | 85.9/11.2/2.9 | 1.78 | 80.7/13.3/6.0 |
| GPT-5.4-mini | 2.1% | 8.77 | 60.5/16.7/22.8 | 4.31 | 85.7/5.4/8.9 |

**Bank analysis.** Tab. 7 summarizes the stage-I banks constructed for EgoWearer identification, where each bank is built from five reference clips. Across builders, extraction produces 2.95–6.92 candidate cues per clip, and the merge process compresses them into 8.2–18.4 final entries per bank, corresponding to compression ratios of 0.53–0.62. The final banks remain organized across all three evidence categories: appearance, owned objects, and behavior. The bank construction is not a simple append-only process. Revision operations account for 25–38% of merge decisions, and 18–27% of final entries are updated through later confirmation or revision, indicating that the bank accumulates and refines reusable personal evidence across clips.

Tab. 8 analyzes how the constructed bank is accessed during stage-II querying. For Gemini-3-Flash [27] and Gemma-4-31B [22], the model requests stored visual evidence for nearly all queries, while still selecting a small subset of the bank: 3.63 and 2.86 requested entries on average, respectively. The model-reported decisive set is smaller, averaging 2.67 and 1.78 entries. The requested and decisive distributions span appearance, owned-object, and behavioral evidence, showing that the bank supports decisions through multiple forms of personal evidence rather than a single cue type. This distribution closely mirrors human reasoning. We manually label a 90-query subset as a reference point, where human accuracy reaches 84.1% (excluding ambiguous cases). The human evaluators cited personal objects (46.3%), appearance (41.5%), and behavior (12.2%) as their primary evidence, validating

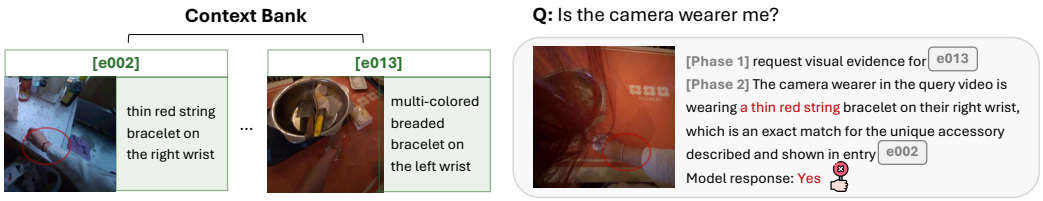


Figure 6: Failure case of our Agentic Context Bank. The model requests visual evidence for entry e013, but does not request the raw visual evidence for entry e002, whose text descriptor later drives the final answer. The query wearer has a different bracelet from the one shown in e002; without revisiting the corresponding visual evidence, the model incorrectly treats the descriptor as an exact match and answers Yes.

our structural choice to organize the bank around multiple complementary cues. GPT-5.4-mini [57] follows a different access pattern, usually deciding from the text view and requesting visual evidence in 2.1% of queries; when it does request evidence, it inspects a larger set. This strong preference for text-based resolution likely explains why its overall performance gain is smaller compared to the other two backbones in Tab. 3. Together, these statistics show that our stage I effectively compresses and refines multi-category personal evidence, and stage II selects a targeted evidence subset for each query.

**Failure cases.** Fig. 6 illustrates a failure mode of the query-adaptive stage. Although the relevant information exists in the constructed bank, the model fails to request the specific visual evidence required to verify a fine-grained detail. Instead, after retrieving a different bracelet-related entry, it relies solely on the text description of e002 and incorrectly infers a match. Because the query frame reveals that the bracelet is visually distinct, the correct answer is No. This example highlights that personalization errors stem from inadequate evidence selection at query time.

## E Broader Impacts

This work studies a capability needed for future personalized multimodal assistants: using visual context from a user's own history to answer user-specific questions. If developed responsibly, Personal VCL could make assistants more helpful in everyday settings. It could support remembering personal objects, recognizing familiar people with user consent, comparing a user's current action to their usual routine, or providing individualized feedback in domains such as cooking, rehabilitation, sports, and craft practice. These applications could be especially valuable when generic visual understanding is insufficient because the relevant information is personal rather than part of public world knowledge.

The same capability also creates risks. Personal visual histories are inherently sensitive. They may encode the identity and behavior of the wearer, the presence and actions of bystanders, private spaces, social relationships, and repeated routines. Models that can organize and reason over this information could be used for surveillance, unwanted identification, behavioral profiling, or continuous authentication without meaningful consent. Errors in such systems could also cause harm if they lead to incorrect personalized guidance or misidentify people, objects, or deviations from routine.

Our work is intended as an evaluation benchmark and inference-time study, not as a deployed personal assistant. We use existing research datasets and focus on measuring whether LMMs can use curated personal visual context. Future real-world Personal VCL systems should require informed consent, clear data ownership, strong access control, secure and privacy-preserving storage, and limits on secondary use of personal visual data. Deployment should also consider bystander privacy and provide mechanisms for deletion, auditing, and user control over what visual memory is retained.