EgoMemory: Memory-Augmented Personalized Retrieval for Long-Context Egocentric Video

Yuanmin Tang[1,2*], Jue Zhang[3], Xiaoting Qin[3], Jing Yu[4], Meikang Qiu[5], Gaopeng Gou[1], Gang Xiong[1], Qingwei Lin[3], Saravan Rajmohan[3], Dongmei Zhang[3], Qi Wu[6]

[1]Institute of Information Engineering, Chinese Academy of Sciences,
[2]University of Chinese Academy of Sciences,
[3]Microsoft, [4]Minzu University of China, [5] University of Augusta, [6]University of Adelaide

Abstract

Recent advances in AI and wearable devices, such as augmented-reality glasses, have made it possible to augment human memory by retrieving personal experiences in response to natural language queries. However, existing egocentric video datasets fall short in supporting the personalization and long-context reasoning required for episodic memory retrieval. To address these limitations, we introduce EgoMemory, a benchmark derived from Ego4D, enriched with 165,795 user-specific object annotations over 245 videos from 45 participants, yielding 639 distinct, human-curated, and evaluated queries for rich and individualized episodic memory retrieval. Leveraging this resource, we present EgoRetriever, a novel, training-free retrieval framework that combines Multimodal Large Language Models with reflective Chain-of-Thought prompting. Our approach enables interpretive inference of user intent and generates detailed target video descriptions by leveraging contextualized personal memory for video retrieval. Extensive experiments on three benchmarks, including EgoMemory, EgoCVR, and EgoLife, demonstrate that EgoRetriever consistently and substantially outperforms state-of-the-art baselines, highlighting its strong generalizability and practical potential for personalized, long-context egocentric video retrieval.

1 Introduction

The integration of AI into wearable technologies (*e.g.*, glasses), suggests a future where human memory is augmented through continuous experience capture and retrieval. This notion closely resembles Vannevar Bush's "Memex", proposed in 1945 as a conceptual system for amplifying cognition through personalized, associative information access (Bush et al., 1945). Recent advances in wearable devices and large language models (LLMs) may bring this long-standing vision within reach.

Central to realizing this vision is the task of episodic memory retrieval (Grauman et al., 2022), which aims to extract relevant visual episodes from a user's egocentric video archives based on natural language queries. Distinct from traditional text-to-video retrieval, this task uniquely emphasizes *personalization* and *long-context*: (i) data are continuously recorded from the user's viewpoint (*i.e., egocentric*); (ii) most queries explicitly reference personal objects (*e.g.*, our empirical analysis in Section 3.2.1 indicates that **88.4%** of queries in the Ego4D dataset (Grauman et al., 2022) exhibit such explicit referencing); (iii) user queries frequently involve specific objects or actions in remote history (*e.g.*, "what is the location I play with my dog in *last month*?"), necessitating solutions capable of long-context video understanding. Current episodic memory retrieval tasks, however, predominantly concentrate on single-video or short-term scenarios (Grauman et al., 2022; Hummel et al., 2024), neglecting the personalized, long-context nature intrinsic to episodic memory retrieval.

To address this limitation, our study focuses explicitly on long-context personal egocentric memory retrieval. Given the absence of explicit annotations for personally relevant objects in existing egocentric video datasets (Singh et al., 2016; Grauman et al., 2022; Hummel et al., 2024; Yang et al., 2025), we first introduce the **EgoMemory** benchmark, designed explicitly for extracting personalized information from users' historical videos to facilitate long-context episodic video retrieval (details in Section 3.2.1). Figure 1(a) exemplifies this by demonstrating how attributes of a user's personal item (a "dog") can be systematically extracted from past video clips and corresponding captions via MLLMs. We apply this pipeline to annotate 245 videos from 45 unique participants in the Ego4D dataset (Grauman et al., 2022), resulting in 165,795 user-specific object annotations to constitute a comprehensive personal memory bank.

[*]Work is done during an internship at Microsoft

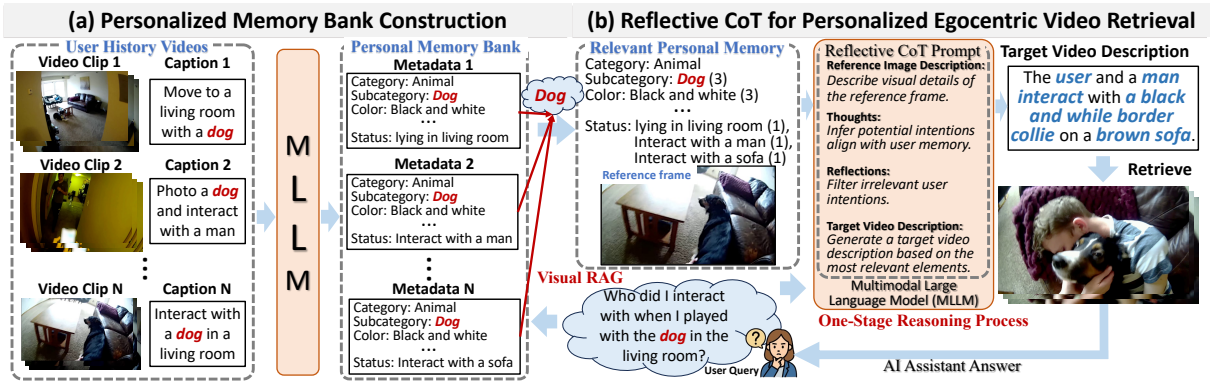

Figure 1: Overview of our approach for personalized egocentric video retrieval, comprising two progressive modules: (a) offline construct a personalized memory bank from each user's historical videos; (b) online retrieval by query-relevant personal memory to guide intention understanding.

Candidate queries in Ego4D are also filtered for personalization and long-context via an MLLM-assisted procedure with final human verification.

With the constructed EgoMemory benchmark, we further propose **EgoRetriever**, a novel, training-free framework tailored explicitly for long-context episodic video retrieval. As shown in Figure 1(b), EgoRetriever combines MLLMs with a reflective Chain-of-Thought (CoT) prompting strategy that performs intent inference and constraint verification over a structured personal memory bank to interpret nuanced user intentions and generate detailed target descriptions for video retrieval. This approach improves accuracy by grounding generation in history-validated, fine-grained cues (e.g., the dog's color) and contextually relevant elements (*e.g.,* "sofa" and "interaction with a man") drawn from personal memory.

To summarize, our main contributions are: (1) We formalize memory-augmented personalized long-context egocentric video retrieval and introduce the **EgoMemory** benchmark, where each query is paired with an individualized memory bank constructed from user-specific object-centric annotations and cross-video recurrence statistics in Ego4D (Grauman et al., 2022). (2) We propose EgoRetriever, a training-free retrieval framework that combines Multimodal Large Language Models (MLLMs) with reflective Chain-of-Thought (CoT) prompting to interpret user queries by leveraging personal memory and generate detailed descriptions for video retrieval. (3) Extensive experiments on EgoMemory, EgoCVR (Hummel et al., 2024), and EgoLifeQA (Yang et al., 2025) show that EgoRetriever consistently outperforms strong baselines, highlighting its generalizability and potential for practical lifelog retrieval.

2 Related Work

Egocentric Datasets and Benchmarks. Early egocentric studies used ADL (Pan et al., 2022), CharadesEgo (Sigurdsson et al., 2018), and EGTEA Gaze+ (Li et al., 2018), but these were limited in scale and diversity. Larger datasets (*i.e.,* EPIC-KITCHENS (Damen et al., 2020) and Ego4D (Grauman et al., 2022)) broadened the field and enabled many tasks. Specialized corpora, including EgoProceL (Bansal et al., 2022), IndustReal (Schoonbeek et al., 2024), HoloAssist (Wang et al., 2023a), EgoExo4D (Grauman et al., 2024), and EgoExoLearn (Huang et al., 2024), target procedural and multi-view understanding. Recent benchmarks such as EgoSchema (Mangalam et al., 2023) and EgoPlan-Bench (Li et al., 2024) emphasize temporal reasoning and planning, while EgoMemoria (Ye et al., 2025) and EgoLife (Yang et al., 2025) provide week-long, multi-participant data for studying longer-term behavior. While these benchmarks emphasize procedural understanding, temporal reasoning, or week-scale daily life, EgoMemory is *complementary*: it specifically targets fine-grained, person-specific variability for long-context *personalized* retrieval, and is, to our knowledge, the first benchmark constructed around per-user object-centric memory banks.

Composed Image and Video Retrieval. Composed image retrieval (CIR) retrieves images that are semantically edited by textual prompts (Vo et al., 2019; Baldrati et al., 2022). Zero-shot CIR methods (Saito et al., 2023; Baldrati et al., 2023; Tang et al., 2024c; Gu et al., 2024; Karthik et al., 2024; Tang et al., 2024b; Suo et al., 2024; Du et al., 2024; Tang et al., 2024a, 2025a) use multimodal encoders such as CLIP (Radford et al., 2021) to

reduce annotation needs, yet often struggle with implicit human intent. Recent training-free approaches (e.g., CIReVL (Karthik et al., 2024) and OSrCIR (Tang et al., 2025b)) leverage large language models to infer intent and improve compositional reasoning without supervision. Extending to video, composed video retrieval addresses temporal complexity. EgoCVR (Hummel et al., 2024) supports fine-grained egocentric queries with a two-stage caption fusion pipeline. Despite progress, current frameworks are still under a model dynamic context and personal relevance in real egocentric scenarios. We introduce a training-free, one-stage retrieval framework that grounds user queries in a dynamic personal memory bank and produces fine-grained video descriptions. This design achieves state-of-the-art performance on EgoMemory and advances personal memory retrieval.

Memory-Augmented Long-Context Retrieval. Retrieval-Augmented Generation (RAG) combines large language models with external memory to enable long-context reasoning (Lewis et al., 2020; Jiang et al., 2023; Shi et al., 2023; Ram et al., 2023; Izacard et al., 2022). Graph-augmented retrieval further supports multi-hop reasoning by leveraging structured knowledge graphs for re-ranking and contextual linking (Ding et al., 2019; Zhu et al., 2021; Nie et al., 2019; Das et al., 2019; Asai et al., 2020; Li et al., 2021), e.g., HippoRAG (Gutiérrez et al., 2024). Recent work extends retrieval augmentation to long videos: VideoAgent (Fan et al., 2024) uses an agentic pipeline to iteratively retrieve evidence, while VideoRAG (Luo et al., 2025) proposes visually aligned retrieval augmentation for long-video comprehension. Lifelogging systems (Rossetto et al., 2020; Nguyen et al., 2021) organize personal data with multimodal knowledge graphs, but often rely on static schemas and offer limited flexibility for user-specific interpretation. In contrast, EgoMemory builds personalized, object-centric memory banks from egocentric video, and EgoRetriever is designed to output a retrieval-oriented target description (rather than a free-form answer) for ranking candidates under long-context personalization.

3 Methodology

3.1 The EgoMemory Benchmark

Although Ego4D NLQ (Grauman et al., 2022) is large-scale, it targets temporal localization within *isolated* clips and neither aggregates videos by user identity nor explicitly captures ownership/user-linkage cues. As a result, direct NLQ evaluation can favor generic scene or frequency priors, and most queries do not require long-horizon reasoning. To bridge these gaps, we introduce EgoMemory, which aggregates full user histories to enforce long-context, personalized retrieval.

Benchmark Construction & Filtering. We treat each user as a distinct retrieval unit, aggregating *all* their videos as personal context. From 137 participants, we manually select 45 with sufficient temporal coverage to simulate realistic AR usage across diverse daily activities. To ensure evaluation probes real personalization rather than scene priors, we apply a rigorous two-stage filtering pipeline: (i) **GPT-4o Pre-screening**: We retain queries with explicit personal references (*e.g.,* possessives, deictic cues) or strong user linkage, discarding generic queries designed for short single-clip answers. *GPT-4o is used only for filtering and metadata generation; the target clips and query–target pairs are inherited from the original NLQ annotations.* (ii) **Human Verification**: We verify that the referenced object is plausibly user-linked across history. Concretely, for each retained query–target pair, annotators review 20 additional short clips from the same user containing the same object class; queries are labeled *personal* when $\geq 90\%$ of reviewed instances match the target and *uncertain* when $\geq 75\%$. This process yields 639 high-quality personal queries across 245 videos ($\sim$91.6% personal).

Dataset Composition. The benchmark includes a candidate pool of 2,228 clips from 45 users (64.25h total), with per-user candidate sizes ranging from 9 to 61 clips and clip durations spanning 4–300 seconds. To reflect realistic lifelog histories that are large and noisy, we provide 165,795 user-specific object annotations across 12 attributes (*e.g.,* category, color, brand) as the *benchmark memory* available to retrieval systems. Our method can either consume this structured memory directly or build an equivalent memory automatically from videos (Sec. 3.2.1). Further details (*i.e.,* distribution statistics in Fig. 10) are provided in Appendix B.1.

3.2 The EgoRetriever Framework

Problem Formulation. We consider a continuous setting where users record egocentric videos $\mathcal{V} = \{V^{(1)}, \ldots, V^{(N)}\}$, segmented into candidate clips $\mathcal{C} = \{C_1, \ldots, C_M\}$. To enable personalization, we construct a user-specific memory bank $\mathcal{M}$ using a pretrained MLLM Ψ_M. The retrieval task

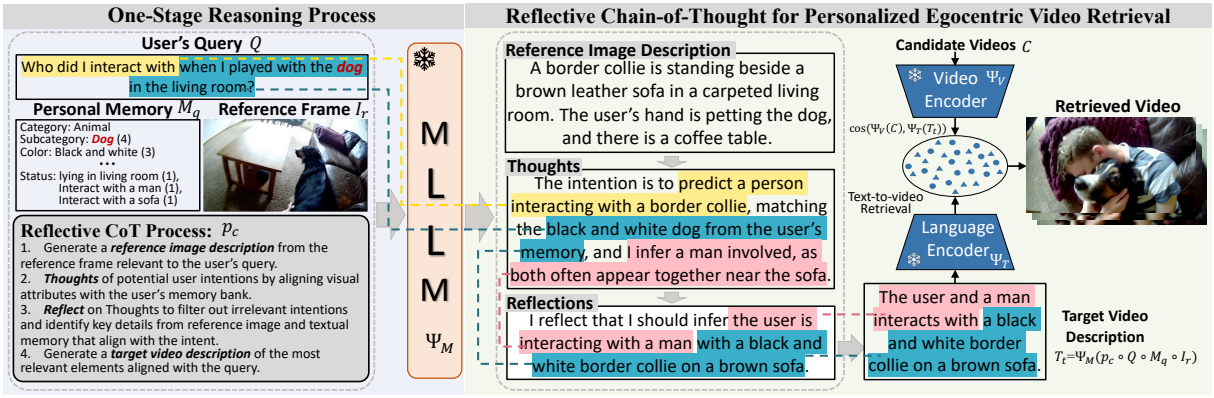

Figure 2: Overview of EgoRetriever. An MLLM processes textual personal memory data M_q, the reference frame I_r and the user's query Q to generate the desired target video description T_t by reflective CoT. A vision-language model is to perform text-to-video retrieval. Texts with different colors show the reasoning traces of each intention.

is defined as: given a natural language query Q, retrieve the most relevant clip $C^* \in \mathcal{C}$. EgoRetriever first queries $\mathcal{M}$ to obtain personal object metadata $\mathcal{M}_q$ and a lightweight visual anchor I_r. Using $(Q, \mathcal{M}_q, I_r)$, the framework employs an MLLM with reflective Chain-of-Thought (CoT) prompting to generate a target description T_t. Finally, a video-language retriever embeds T_t (via text encoder Ψ_T) and each candidate C_i (via video encoder Ψ_V) to rank clips by cosine similarity:

$$C^* = \underset{C_i \in \mathcal{C}}{\texttt{argmax}} \frac{\Psi_V(C_i)^\intercal \Psi_T(T_t)}{\|\Psi_V(C_i)\| \|\Psi_T(T_t)\|}. \quad (1)$$

3.2.1 Personalized Memory Bank

To prioritize user context over incidental scene exposure, we note that **88.4%** of Ego4D (Grauman et al., 2022) queries target physical objects, estimated via lightweight query analysis with SpaCy (Honnibal et al., 2020) and WordNet (Miller, 1995). We operationalize personalization as *personally experienced objects*, supported by cross-video recurrence (Lee et al., 2012; Yang et al., 2025). EgoMemory exposes a large and noisy *benchmark memory* from daily life, while EgoRetriever consolidates it into a structured memory bank that retains user-linked, history-consistent cues for long-context retrieval (Figure 1). EgoRetriever uses this consolidated memory at inference to down-weight one-off co-occurrences.

Construction. We represent each user's history as a structured memory $\mathcal{M}$ of object-centric entries (attributes and recurrence cues). In EgoMemory, $\mathcal{M}$ is instantiated from the provided object annotations; more generally, the same schema can be populated from videos and narrations using a pre-trained MLLM Ψ_M (Appendix Figure 11). Compared to unstructured transcripts (Yang et al., 2025), this design makes user-linkage explicit and reduces sensitivity to background scene noise.

Visual Retrieval-Augmented Generation. At inference, we first parse the query Q with SpaCy (Honnibal et al., 2020) to extract its object-centric entity (*e.g.,* "drawstring bag"), which serves as the key for retrieving a small subset of memory entries $\mathcal{M}_q \subset \mathcal{M}$ via semantic matching, filtered by recurrence so that frequently observed, user-linked objects are prioritized. The retrieved entries are then aggregated into a frequency-weighted attribute summary M_q (per-attribute mode/frequency over the entry set), which is concatenated into the prompt as compact textual context. We further select a lightweight visual anchor I_r from the same retrieved context: I_r is the middle frame of a *retrieved historical reference clip* V_r (the centroid of the retrieved memory entries) and is *not* a frame from the ground-truth target clip C^*, ruling out target leakage (ablation in Appendix Table 9). The MLLM conditions on $(Q, \mathcal{M}_q, I_r)$ to generate a retrieval-oriented target description, effectively performing *retrieve-then-generate* over personal memory and reducing reliance on generic scene priors.

3.2.2 Reflective Chain-of-Thought

Prior egocentric retrieval methods (e.g., TFR-CVR) often adopt a two-stage pipeline that captions the reference frame and then rewrites it with an LLM, which can discard fine-grained visual cues and user-specific details. EgoRetriever instead uses a one-stage, training-free design that prompts an MLLM to directly generate the target video description T_t. Formally, given an MLLM Ψ_M, we define:

$$T_t = \Psi_M(p_c \oplus Q \oplus \mathcal{M}_q \oplus I_r), \quad (2)$$

where $\oplus$ denotes concatenation and p_c is our reflective CoT prompt. As shown in Figure 2, the process has four stages:

Reference Image Description. First, the MLLM generates a dense semantic description of the visual anchor I_r, prompted to preserve personalization-relevant cues (*e.g.,* object states and surrounding context) while suppressing irrelevant background noise. For example, in Figure 2, it isolates the "border collie" and "leather sofa" as key anchors.

Thoughts (Intent Reasoning). Conditioned on the query Q and the visual description, the model interprets the user's implicit intent. It explicitly correlates visual attributes from I_r with user-specific patterns retrieved from the memory bank $\mathcal{M}_q$. This step hypothesizes potential interactions, effectively bridging the gap between the static reference frame and the dynamic target event.

Reflections (Constraint Checking). To mitigate the hallucination issues common in MLLMs, this stage acts as a self-verification mechanism. The model critiques its own "Thoughts" against the hard constraints of the visual evidence I_r and the retrieved memory $\mathcal{M}_q$. Implausible assumptions (*e.g.,* interactions inconsistent with the user's history) are filtered out, ensuring the reasoning trajectory remains grounded.

Target Video Description. Finally, the MLLM synthesizes the verified reasoning traces into a focused target description T_t. This description is optimized for retrieval, abstracting the user's intent into a search query that captures the likely visual appearance of the target clip (*e.g.,* describing the action of "petting the dog").

Retrieval. The generated description T_t is then used to rank the candidate set $\mathcal{C}$ via cosine similarity in the embedding space of a pretrained video-language model (*e.g.,* EgoVLPv2) as in eq.1. Further details are provided in Appendix A.3.1.

4 Experiments

Evaluation Metrics. We adopt **mean Recall@K** across users as our principal evaluation metric, reporting mean Recall@1, mean Recall@2, and mean Recall@3. Specifically, for each user, we compute Recall@K based on their individual candidate set and then average across all users to obtain a macro-level performance summary. This approach ensures fair contribution from each user, mitigating the bias that could arise from varying query counts per user. Similar evaluation metrics are also adopted in Ego4D Episodic Memory benchmarks (Grauman et al., 2022) and EgoCVR (Hummel et al., 2024). Moreover, in settings with a single answer per query, Recall@K is equivalent to Hit Rate@K, widely accepted in recommender systems (Sun et al., 2019). Candidate set statistics are provided in the Appendix B.2.

Implementation Details. We use GPT-4o to construct user-specific memory banks by generating object-centric metadata from video clips, and also for reflective CoT reasoning. Retrieval experiments were run on four NVIDIA V100 GPUs (32GB). We evaluate several video-language models, including LanguageBind (Zhu et al., 2024), CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and EgoVLPv2 (Pramanick et al., 2023), and employ EgoVLPv2 as the text encoder in EgoRetriever. CLIP/BLIP video features are computed by averaging embeddings over 15 uniformly sampled frames; candidate videos are matched with GPT-4o-generated textual descriptions via cosine similarity. Annotating all 165,795 object attributes costs $\sim$ \$670 via GPT APIs (one-time); after construction, no additional API calls are required for memory-bank updates beyond processing newly added clips. More details are in Appendix F.1.

Benchmarks and Baselines. We evaluate EgoRetriever on three benchmarks to assess personalization and generalization: (i) **EgoMemory** for personalized retrieval requiring long-horizon context; (ii) **EgoCVR** (Hummel et al., 2024) for composed video retrieval under *global search* (ranking over the full corpus) and *local search* (ranking over a temporally restricted candidate set); and (iii) **EgoLifeQA** (Yang et al., 2025) for long-context episodic QA, covering all five official categories: *EntityLog*, *EventRecall*, *HabitInsight*, *RelationMap*, and *TaskMaster*.

We compare against three categories of SoTA systems: (i) training-free encoders with late fusion/averaging (CLIP, EgoVLPv2, and LanguageBind, a unified video–language embedding model); (ii) composed retrieval methods (*e.g.,* CIReVL, OSrCIR); and (iii) egocentric long-context captioning/reasoning systems, including **TFR-CVR** (Hummel et al., 2024), **VideoAgent** (Fan et al., 2024), and **VideoRAG** (Luo et al., 2025) (a visually-aligned retrieval-augmented baseline). For EgoLifeQA, we additionally include EgoGPT (specialized lifelog QA) and LLaVA-OV (generalist MLLM). For fairness, all MLLM-invoking methods (EgoRetriever, TFR-

Method	Video Model	Textual Memory	Visual Ref	Fusion Strategy	Mean Recall (%) mR@1	mR@2	mR@3
Random	✗	✗	✗	—	3.62	9.74	15.23
CLIP (Radford et al., 2021)	✗	✓	✓	Avg	15.64	18.63	22.71
BLIP (Li et al., 2023)	✗	✓	✓	Avg	16.02	19.17	24.12
EgoVLPv2 (Pramanick et al., 2023)	✓	✓	✓	Avg	16.67	21.71	25.09
LanguageBind (Zhu et al., 2024)	✓	✓	✓	Avg	16.16	21.24	24.34
$\text{BLIP}_{\text{CoVR}}$ (Ventura et al., 2024)	✗	✓	✓	Cross-Attn.	15.94	19.17	23.00
$\text{BLIP}_{\text{CoVR-ECDE}}$ (Thawakar et al., 2024)	✗	✓	✓	Cross-Attn.	16.41	19.63	23.64
CIReVL (Karthik et al., 2024)	✗	✓	✓	Captioning	16.95	20.13	24.37
OSrCIR (Tang et al., 2025b)	✗	✓	✓	Captioning	17.28	21.64	25.49
VideoAgent (Fan et al., 2024)	✓	✓	✓	Captioning	17.49	26.40	35.62
TFR-CVR (Hummel et al., 2024)	✓	✓	✓	Captioning	18.21	27.12	32.05
VideoRAG (Luo et al., 2025)	✓	✓	✓	Captioning	<u>19.70</u>	<u>30.83</u>	<u>39.82</u>
EgoRetriever (Ours)	✓	✓	✓	Captioning	**23.19**	**38.48**	**47.83**

Table 1: Mean Recall@K (%) on the EgoMemory benchmark. "Textual Memory" denotes a personal text-based memory bank, and "Visual Ref" denotes the visual anchor. The complete results table is provided in Table 7.

EgoCVR (Composed Video Retrieval) (Hummel et al., 2024)

Method	Global Search R@1	R@5	R@10	Local Search R@1	R@2	R@3
CIReVL	2.0	6.8	10.6	33.6	49.7	61.4
OSrCIR	4.9	9.3	13.4	37.4	53.3	68.1
TFR-CVR	14.7	41.2	55.6	46.1	62.4	73.9
EgoRetriever	**17.4**	**49.2**	**62.7**	**50.3**	**68.2**	**76.4**

EgoLifeQA (Episodic QA) (Yang et al., 2025)

Method	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Avg
GPT-4o	34.4	42.1	29.5	30.4	44.4	36.2
LLaVA-OV	36.8	34.9	31.1	–	–	–
EgoGPT	39.2	36.5	31.1	33.6	39.7	36.0
EgoRetriever	**42.5**	**45.1**	**37.7**	**35.8**	**46.2**	**41.5**

Table 2: Comprehensive generalization analysis on EgoCVR and EgoLifeQA benchmarks.

CVR, VideoAgent, VideoRAG) use the same GPT-4o backbone. Details are in Appendix D.1 and D.2.

4.1 Quantitative and Qualitative Results

Performance on EgoMemory. In Table 1, we report mean Recall@K on EgoMemory, which stresses personalized retrieval under long-horizon context and intent-critical temporal relations. EgoRetriever achieves 23.19% mR@1 and 47.83% mR@3, yielding the best overall performance among all compared methods. VideoRAG (Luo et al., 2025) is the strongest baseline on this benchmark, yet EgoRetriever still improves over it by 6.38% on average. In addition, EgoRetriever improves over TFR-CVR and VideoAgent by 4.98% and 5.70% in mR@1, and exceeds the strongest CIR baseline OSrCIR by 5.91% in mR@1. Overall, while retrieval-augmented long-video systems benefit from strong evidence retrieval, they can still retrieve visually or lexically similar but intent-inconsistent moments. In contrast, our reflective CoT prompting combined with a structured personal memory bank better aligns retrieval with user-specific, history-dependent constraints, leading to consistent performance gains. Additional results are provided in Appendix Table 7.

Generalization Analysis. We further evaluate robustness on EgoCVR and EgoLifeQA (Table 2). On EgoCVR, EgoRetriever consistently outperforms the two-stage baseline TFR-CVR in both Global and Local settings, with an average gain of 5.05% computed as the unweighted mean over the six reported metrics (R@1/5/10 for Global and R@1/2/3 for Local). On EgoLifeQA, EgoRetriever outperforms EgoGPT by an average of +5.5% across all five categories (EntityLog +3.3, EventRecall +8.6, HabitInsight +6.6, RelationMap +2.2, TaskMaster +6.5). Collectively, these results indicate that the same training-free pipeline generalizes beyond EgoMemory to composed video retrieval and long-context episodic QA, benefiting from memory structuring and reflective reasoning rather than task-specific training.

Qualitative Analysis. We further analyze representative examples in Appendix Figure 13 to illustrate the reasoning behavior of EgoRetriever. The visualizations show that the *Reflective* stage is crucial for filtering irrelevant visual clutter and correcting initial reasoning errors (*e.g.,* misinterpreting *pre- vs. post-action* constraints such as "before I dropped them", or overlooking intent-critical interaction cues). Rather than relying on dominant scene priors, EgoRetriever explicitly validates the decisive constraints against the structured personal memory, leading to history-consistent retrievals aligned with the user's intent. Quantitative compar-

	Methods	mR@1	mR@2	mR@3
1.	Full Model (Reflective CoT)	23.19	38.48	47.83
Significance of key modules				
2.	w/o Textual Memory	18.04	30.73	36.91
3.	w/o Reference Frame	20.79	32.94	40.23
4.	w/o Original Description	21.49	36.04	43.70
5.	w/o Thoughts	20.14	33.62	41.89
6.	w/o Reflection	20.52	35.17	42.90
7.	w/o ICT	21.37	36.29	43.49
Practical feasibility				
10.	Human captions	23.19	38.48	47.83
11.	EgoGPT auto-captions	22.73	36.72	46.02
12.	GPT-4o auto-captions	21.19	35.22	45.39
Impact of different MLLMs				
13.	LLaVA	20.37	33.58	41.90
14.	Qwen2.5-VL	22.03	35.24	45.27
15.	GPT-4o-mini	22.31	37.19	46.43

Table 3: Ablation results demonstrating the necessity of the key modules in EgoRetriever.

isons against all baselines, including VideoRAG, are reported in Table 1.

4.2 Ablation Study and Performance Analysis

In Table 3, we assess the contribution of each component on EgoMemory. **(1) Models '2–7' evaluate the necessity of key modules within EgoRetriever.** Removing textual memory yields the largest drop in mean Recall (model '2') by 7.94% compared to our full model ('1'), underscoring the need for user-linked metadata. Similarly, the absence of the reference frame (model '3') leads to a 5.18% drop, emphasizing its critical role in grounding the visual context. Within the reflective CoT, omitting original description, thoughts, or reflection reduces performance by roughly 3%–5%, and removing ICT examples gives a smaller but consistent decline. Together, these results show that memory and a lightweight visual anchor carry most of the gain, while each CoT step contributes additive improvements. **(2) Reflective CoT vs. alternative CoT prompting.** Replacing our reflective CoT with simple CoT or an advanced two-stage CoT (DDCoT) degrades mean Recall by about 5% (detailed in Appendix Table 9), indicating the advantage of single-prompt reflective reasoning for interpreting multimodal user intent. **(3) Models '10–12' examine the practical feasibility without human-written narrations.** We re-annotated all 165,795 objects using EgoGPT and GPT-4o captions for each reference frames. Compared to human captions (model '10'), EgoGPT auto captions (model '11') and GPT-4o auto captions

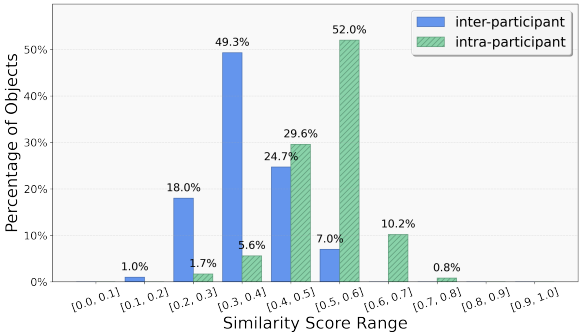

Figure 3: Similarity of top 100 objects in constructed memory banks. The high intra-participant similarity highlights the personalized nature of our benchmark.

(model '12') show only minor declines, confirming that EgoRetriever remains effective without ground truth narrations and is practical for real-world application.**(4) Models '13–15' examine the impact of different MLLMs on performance.** Utilizing open-source MLLMs such as LLaVA (Liu et al., 2023) (model 13') and Qwen2.5-VL (Yang et al., 2024) (model 14') achieves competitive but clearly inferior results compared to GPT-4o, with performance gaps of 4.55% and 2.32%, respectively. Notably, GPT4o-mini (model '15') performs closely to GPT-4o, with only a minor decline of 1.19%, indicating that GPT4o-mini offers a balance between efficiency and retrieval performance.

4.3 Analysis

In this subsection, we provide detailed analyses of our design choices and the common failure cases.

Personalization validation by cross-user memory swap. To verify that the gains of EgoRetriever stem from *user-linked* memory rather than from generic background priors, we swap each user's memory bank with that of a different user at inference time, keeping the query and candidate pool unchanged. Table 4 reports retrieval performance under four conditions: *Matched* (the user's own memory; identical to row 1 of Table 3), *Random-user swap* (memory drawn from a random other user), *Far-user swap* (memory from the most dissimilar user by Jaccard distance over object attributes), and *w/o Memory* (identical to row 2 of Table 3). Both swap conditions degrade sharply below the *w/o Memory* baseline, indicating that mismatched personal memory is actively harmful: the model is not simply benefiting from any generic context but from *the right user's* context. The Far-user swap is consistently worst, confirming that the larger the personalization mismatch, the larger the

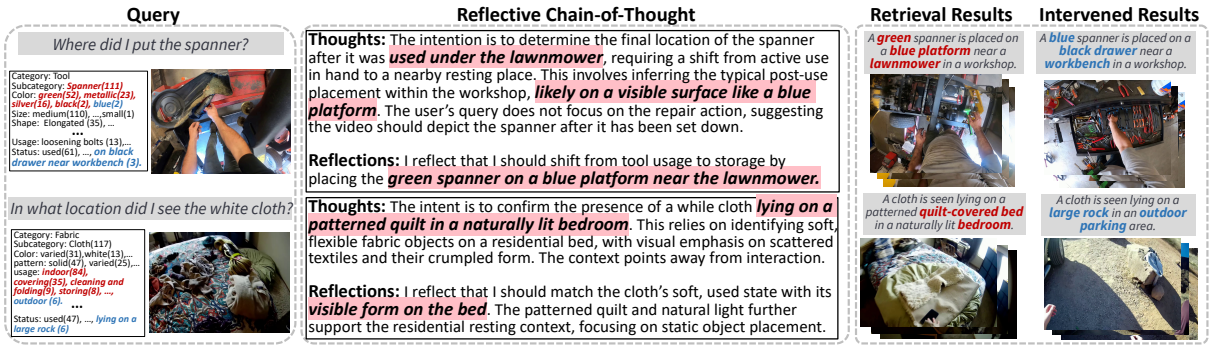

Figure 4: Visualization of common failure cases. The red color denotes the incorrect reasoning outcomes of intention. The top-1 retrieval result and the intervened correction are shown.

Setting	mR@1	mR@2	mR@3
Matched (own memory)	**23.19**	**38.48**	**47.83**
Random-user swap	16.24	27.83	35.09
Far-user swap	14.11	23.38	30.77
w/o Memory	18.04	30.73	36.91

Table 4: Cross-user memory swap on EgoMemory. Mismatched personal memory underperforms even *w/o Memory*, isolating the contribution of user-linked context.

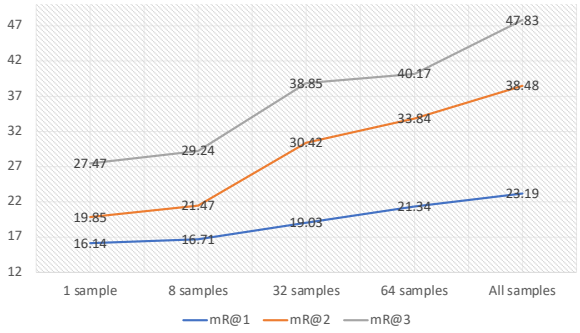

Figure 5: Effect of the number of metadata for each object in the memory bank.

performance loss.

Analysis of Personalization Heterogeneity. To quantify user specificity, we compute the Jaccard similarity over attribute sets for the 100 most frequent object types, comparing inter- vs. intra-participant distributions. Furthermore, to control for environmental bias, we recompute inter-participant similarity by restricting comparisons to matched coarse scenes (kitchen, living room, outdoor) using the "Status" metadata, formulated as $J(A^{\text{scene}=s}, B^{\text{scene}=s}) = \frac{|A \cap B|}{|A \cup B|}$. As illustrated in Figure 3, **68.3%** of objects exhibit similarity scores below 0.4, indicating that personalization persists significantly beyond scene priors.

Impact of Memory Bank Context Length. We evaluate the influence of context length by varying the volume of retained object metadata (Figure 5). Restricting memory to short-term contexts (1–8 samples) significantly limits performance (mR@3 $< 30\%$), as essential long-term user patterns remain underrepresented. As the context length increases, incorporating a broader history of user experiences, retrieval accuracy improves markedly. Conversely, utilizing the full memory bank yields substantial gains, elevating mR@1 by over 7% and mR@3 by $\sim 20\%$ compared to the single-sample baseline. These findings underscore the importance of a comprehensive and extensive memory bank in enabling accurate long-context video retrieval.

Analysis of Common Failure Cases. We analyze 100 failures and find two dominant error types (Fig. 4): *object disambiguation* (74%), where the model confuses visually similar targets in clutter (*e.g.,* selecting the wrong spanner in a workshop; Row 1), and *context misinterpretation* (21%), where ambiguous references lead to incorrect scene grounding (*e.g.,* retrieving an indoor "white cloth" instead of the correct outdoor context; Row 2). Adding explicit contextual cues (e.g., "black drawer," "outdoor") often reduces these errors, suggesting that stronger object differentiation and context reasoning remain key limitations for personalized long-horizon egocentric retrieval.

Efficiency Analysis. Table 5 compares query-time latency and API cost against retrieval quality on EgoMemory. EgoRetriever offers the best overall accuracy–efficiency trade-off. Relative to TFR-CVR (GPT-4o), our GPT-4o variant reduces latency from 1.00 s to 0.70 s (30% relative reduction) and lowers per-query API cost from \$0.007 to \$0.004 (43% reduction), while improving average retrieval performance by 6.9%. The GPT-4o-mini variant is the most efficient configuration, achieving 0.50 s latency and \$0.002 per query, yet still outperforming TFR-CVR by 5.6% on average, suggesting EgoRetriever is practical for interactive use.

Growing-memory evaluation. We simulate an online setting by constructing each user's memory bank from the first $20, 40, 60, 80, 100\%$ of videos

Method	Backbone	Lat.(s)	Cost($)	Avg.
CIReVL	GPT-3.5	1.40	**0.001**	24.9
OSrCIR	GPT-4o	0.70	0.004	27.9
TFR-CVR	GPT-4o	1.00	0.007	41.3
EgoRetriever	GPT-4o-mini	**0.50**	0.002	46.9
EgoRetriever	GPT-4o	0.70	0.004	**48.2**

Table 5: Efficiency comparison. Inference latency per query and API cost with average retrieval performance.

Prefix (% videos)	20	40	60	80	100
Avg.	27.46	31.29	33.07	35.61	36.50

Table 6: Growing-memory evaluation on EgoMemory: memory uses the first $p\%$ videos (chronological); only queries with targets in-prefix are evaluated (macro-avg).

in chronological order and evaluating only queries whose targets lie within the available prefix to avoid leakage. As shown in Table 6, average retrieval performance increases monotonically by 9.04% from 20% to 100% as memory accumulates, while latency and API cost remain stable due to offline precomputation (Table 5; Appendix Table 6).

5 Conclusion

In this paper, we address the challenge of personalized, long-context episodic retrieval by introducing EgoMemory, a benchmark enriched with user-specific memory banks. To tackle this, we propose EgoRetriever, a training-free framework leveraging MLLMs and reflective Chain-of-Thought to ground user queries in personal memory explicitly. Extensive experiments demonstrate state-of-the-art performance and strong generalization, advancing personalized long-context egocentric retrieval and inspiring future research on user-centric memory augmentation. We will publicly release the EgoMemory annotations, prompts, evaluation scripts, and EgoRetriever code upon publication.

Limitations

Our study focuses on a training-free retrieval setting that leverages a structured, object-centric memory bank constructed from egocentric videos. As a result, performance is influenced by the coverage and fidelity of the underlying extracted metadata, as well as by the reasoning behavior of the chosen foundation models. In addition, while the benchmark is motivated by long-horizon lifelog retrieval, our evaluation is conducted in an offline setting with fixed user histories and candidate pools; handling fully online streams with continual memory growth, update policies, and resource constraints remains an important direction. Finally, EgoMemory is derived from Ego4D and reflects the distributions and annotations of that source, so broader validation across additional egocentric datasets and user populations is still needed.

Ethical considerations

Our benchmark is derived from Ego4D (Grauman et al., 2022), and any use of the benchmark should follow the original data licenses, consent protocols, and access restrictions. Personalized retrieval from egocentric video can expose sensitive information about users and bystanders, and it may be misused for surveillance or manipulation. Responsible deployment, therefore, requires safeguards beyond the scope of this work, such as strong access control, encryption, data minimization, and clear user-facing transparency on what is stored and retrieved, alongside ongoing fairness and misuse monitoring.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. *arXiv:2303.15247*.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474.

Siddhant Bansal, Chetan Arora, and C.V. Jawahar. 2022. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision (ECCV)*.

Vannevar Bush and 1 others. 1945. As we may think. *The atlantic monthly*, 176(1):101–108.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141.

Rajarshi Das, Ameya Godbole, Dilip Kavarthapu, Zhiyu Gong, Abhishek Singhal, Mo Yu, Xiaoxiao Guo, Tian Gao, Hamed Zamani, Manzil Zaheer, and Andrew

McCallum. 2019. Multi-step entity-centric information retrieval for multi-hop question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 113–118, Hong Kong, China. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703, Florence, Italy. Association for Computational Linguistics.

Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. 2024. Image2sentence based asymmetrical zero-shot composed image retrieval. *arXiv preprint arXiv:2403.01431*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, and 1 others. 2024. 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)*, pages 19383–19400.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. 2024. Language-only efficient training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong Lu, Yali Wang, and 1 others. 2024. 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*.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. 2024. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pages 1–17. Springer.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-language for training-free compositional image retrieval. In *The Twelfth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and Bingquan Liu. 2021. Hopretriever: Retrieve hops over wikipedia to answer complex questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:13279–13287.

Yin Li, Miao Liu, and James M. Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision (ECCV)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2025. Video-rag: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems (NeurIPS)*. Accepted to NeurIPS 2025. Preprint: arXiv:2411.13093.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024a. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024b. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14420–14431.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36:25081–25094.

Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Nguyen Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. 2021. Lifeseeker 3.0: An interactive lifelog search engine for lsc'21. In *Proceedings of the 4th annual on lifelog search challenge*, pages 41–46.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. 2022. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197.

Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. 2022. Youhome system and dataset: Making your home know you better. *IEEE International Symposium on Smart Electronic Systems (IEEE - iSES)*.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernischová, and Abraham Bernstein. 2020. Lifegraph: a knowledge graph for lifelogs. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, pages 13–17.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, and 1 others. 2024. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4365–4374.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *ArXiv*, abs/2301.12652.

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *Preprint*, arXiv:1804.09626.

Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. 2016. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. 2024. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962.

Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, and Qi Wu. 2024a. Manipulation intention understanding for accurate zero-shot composed image retrieval.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gaopeng Gou, Gang Xiong, and Qi Wu. 2024b. Denoise-i2w: Mapping images to denoising words for accurate zero-shot composed image retrieval. *arXiv preprint arXiv:2410.17393*.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. 2025a. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. *arXiv preprint arXiv:2503.17109*.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024c. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5180–5188.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 2025b. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14400–14410.

Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. 2024. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. CoVR: Learning composed video retrieval from web video captions. In *AAAI*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, and 1 others. 2023a. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281.

Ying Wang, Yanlai Yang, and Mengye Ren. 2023b. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, and 1 others. 2025. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. 2025. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *The Thirteenth International Conference on Learning Representations*.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *Preprint*, arXiv:2401.02582.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*.

Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. 2021. Adaptive information seeking for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3626, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Appendix

Appendix Outline

A Prompts, Algorithms, and Reasoning Templates

A.1 Personalized Memory-Bank Construction Prompt

In constructing the personalized memory bank, we leverage a structured prompt that guides a pre-trained multimodal large language model (MLLM) to systematically extract detailed object attributes from textual narrations and visual contexts, as outlined in Figure 6. The prompt explicitly instructs the model to identify and describe the primary object interacted with by the user, ensuring structured output consistency in JSON format. This structured extraction facilitates precise aggregation and retrieval of personalized attributes critical for the memory bank.

Structured JSON Format. The prompt mandates the generation of a strictly structured JSON object with explicitly defined top-level keys (*e.g.,* major category, subcategory, color, texture, shape, material, brand, style, pattern, feature, usage, status). Each key corresponds to a specific attribute dimension necessary for capturing fine-grained personal contextual details, ensuring uniformity and ease of downstream processing.

Specificity and Descriptive Precision. To maximize the accuracy and richness of the memory bank, the prompt instructs the model to provide detailed, descriptive attribute values. It explicitly discourages vague descriptions, advocating specificity, for instance, specifying "dark blue with yellow stripes" rather than a generic label like "blue" or "patterned." This precision enhances the utility of the memory bank for fine-grained retrieval.

Known Values and Inventive Flexibility. The prompt includes dynamically populated examples of known attribute values, providing clarity and consistency in expected responses. However, recognizing the inherent novelty in egocentric video contexts, the prompt encourages the model to introduce new, concise, and descriptive attribute values when existing examples are insufficient, thereby continually enriching the attribute taxonomy.

Handling Uncertainty. To maintain reliability and mitigate incorrect assumptions, the prompt explicitly instructs the model to use "unspecified" or "N/A" when an attribute value cannot be confidently determined from available information. This approach preserves the integrity and trustworthiness of the memory bank by avoiding low-confidence guesses.

Compound Attributes and Consistency. The prompt clearly addresses compound attributes (e.g., combining "fabric and wood" for material attributes), requiring these to be succinctly represented as unified strings. Additionally, it underscores consistency across responses, ensuring that all outputs adhere strictly to valid JSON formatting with appropriate string escaping. This structured consistency facilitates seamless integration into the personalized memory bank infrastructure.

Collectively, these explicit instructions ensure the prompt's effectiveness in systematically extracting detailed, personalized attributes crucial for constructing a robust and reliable personalized memory bank for egocentric video retrieval.

A.2 Algorithm of Calculating Cross-User Object Attribute Similarity

To quantitatively evaluate the attribute diversity of object classes across different users within our proposed personal memory bank, we introduce a systematic method based on attribute similarity metrics. Specifically, we employ the **Jaccard index**, a widely recognized measure for quantifying similarity between finite attribute sets. The discrete and non-hierarchical nature of our object attribute metadata makes the Jaccard index particularly suitable for this analysis.

Algorithm 1 formally describes the calculation procedure. For each object class, the algorithm computes pairwise Jaccard similarity scores between attribute sets associated with every unique pair of users who interact with the same object. Subsequently, the algorithm derives an aggregate diversity score by taking the complement of the average of these pairwise similarities. This aggregate metric, termed the *average diversity*, intuitively captures the heterogeneity in how users characterize identical objects in their personalized memory banks. A higher average diversity score explicitly indicates a richer variability in user-specific object descriptions, underscoring the contribution of our personalized memory bank design toward enhanced contextual representation.

A.3 Details for Our EgoRetriever's Process

A.3.1 Algorithm of EgoRetriever's Process

Algorithm 2 outlines the comprehensive procedure of EgoRetriever for training-free, long-context personal egocentric video retrieval. The process initiates with a natural language query Q, a candidate set of video clips $\mathcal{C}$, and a user-specific semantic memory bank $\mathcal{M}$. EgoRetriever first leverages the MLLM Ψ_M to consult the memory bank $\mathcal{M}$ and retrieve pertinent personal object metadata $\mathcal{M}_q$ and a reference frame I_r. Subsequently, using these contextual cues along with the original query Q and a reflective CoT prompt p_c, a detailed target clip description T_t is generated. This description T_t is then encoded using a text encoder Ψ_T. Each candidate clip $C_i \in \mathcal{C}$ is encoded using a video encoder Ψ_V. The final retrieval of the target clip C^* is achieved by computing the cosine similarity between the encoded description and each encoded candidate clip. This approach allows for a modular retrieval pipeline where the core reasoning and description generation are handled by the MLLM, independent of the specific video-language encoders used, requiring no additional training.

A.3.2 Qualitative Analysis for Personalized Memory Bank

To better understand the practical implications of our proposed personalized memory bank, we conducted a qualitative analysis comparing attribute representations of identical object categories across different users. As an illustrative example in Figure 11, we examined the concept of "dog" as represented by two distinct users (Person A and Person B), each with their own historical interactions captured within their respective personal memory banks.

The detailed attribute annotations reveal substantial variations between the two users in aspects such as color, pattern, style, usage, and specific interaction contexts. Quantitatively, the computed similarity score between these two users' personal memory for the concept of "dog" (Indoor *vs* Outdoor) is notably low, at **0.264**. This underscores a significant divergence in their individual conceptualizations and experiences associated with the same general object class.

The low similarity score highlights a crucial insight: object attributes are perceived and recalled uniquely by different users based on their personal

experiences and contexts. Thus, it clearly demonstrates the necessity and importance of constructing personalized memory banks, as generic or aggregated memory representations would inadequately capture the rich variability in individual user interactions and perceptions. Our findings reinforce the core contribution of our personalized memory bank framework, its ability to accurately reflect nuanced user-specific memory contexts, ultimately enhancing personalized retrieval performance.

A.3.3 Reflective CoT Prompt

A.4 Complete Template for Reflective CoT in Multimodal Video Retrieval

The complete template for our reflective Chain-of-Thought (CoT) reasoning prompt designed for multimodal video retrieval is detailed in Figure 7. This structured prompt systematically integrates visual observations, personalized object attributes, and user query intentions within a unified reasoning framework. Initially, the *Original Image Description* step meticulously documents visual details from the provided reference frame, ensuring the inclusion of all relevant contextual cues. Subsequently, the *Thoughts* step explicitly interprets the user's retrieval intention by analyzing the alignment of visual attributes with personalized object usage information. The subsequent *Reflections* step involves a rigorous evaluation of identified visual and semantic elements, isolating those most congruent with the user's implicit intent. Finally, the *Target Video Description* synthesizes the reflective insights into a succinct, purpose-driven description optimized for accurate retrieval. Importantly, the reflective CoT process is encapsulated within a single comprehensive prompt, promoting coherent, efficient, and interpretable reasoning.

Original Image Description. In this phase, the multimodal large language model (MLLM) is tasked with comprehensively describing *all visible objects and their respective attributes (e.g., color, shape, texture, size)*. Additionally, the model must document *immediate surroundings and broader contextual factors (environmental conditions, indoor/outdoor setting)*, prioritizing precision and detail to preserve critical visual evidence essential for subsequent analytical steps.

Thoughts. Utilizing both the visual description and personalized object attributes (reflecting habitual usage patterns), the MLLM explicitly *interprets the retrieval intent underlying the user's query*. It identifies and elaborates on visual elements (such as dominant colors, textures, or spatial configurations) closely aligning with the user's specific object profile. Further, the MLLM incorporates semantic considerations (such as temporal sequences or action relevance) essential to accurately infer the retrieval context.

Reflections. In this evaluative stage, the MLLM reexamines the highlighted visual and personal object attributes from prior steps. The model critically *summarizes the integration of these visual and usage details in informing its retrieval decision*. It explicitly highlights pivotal elements (e.g., distinguishing material characteristics, contextual environment) and articulates meta-reasoning justifications to ensure coherence between the reference imagery, object attributes summary, and the user's retrieval intention. It reflects precisely on the visual or usage-derived cues that underpin its decision-making rationale.

Target Video Description. Utilizing refined insights from the reflective analysis, the MLLM generates a concise and targeted description pinpointing the specific video segment containing the queried object or interaction. This description is explicitly formulated as a *single, precise sentence* encompassing only retrieval-relevant elements, thus facilitating efficient and highly accurate retrieval.

A.5 Vision-by-Language In-Context Learning Details

Effectively executing Reflective Chain-of-Thought (CoT) reasoning in multimodal large language models (MLLMs) requires not just general instructions but also concrete demonstrations of the reasoning process. To achieve this under a zero-shot setting without relying on direct visual guidance, we adopt a vision-by-language in-context learning (ICL) strategy inspired by recent advances in multimodal reasoning methodologies (Wei et al., 2022; Mitra et al., 2024a; Zheng et al., 2023; Tang et al., 2025b).

Our Reflective CoT ICL provides MLLMs with structured language-based exemplars that guide the model through each reasoning step solely via textual information. As depicted in Figure 8, each example comprises clearly delineated components: an *Original Image Description*, *Thoughts*, *Reflections*, and a *Target Video Description*.

For clarity, consider the following example based on a user query and a provided object attributes summary:

User Query: *"Where was the dog after I laid the bed?"*

Visual Reference: *Middle frame from a reference video that shows a bed's large rectangular form with a decorative patterned cover.*

Object Attributes Summary: Detailed semantic attributes related to the bed, including aspects such as "decorative patterned fabric," "large size," and "used and slightly messy" status.

The Reflective CoT steps are as follows:

- **Original Image Description:** The MLLM generates a detailed depiction of visually pertinent components to the user query. In this scenario, the description captures the bed's large rectangular form with a decorative patterned cover, noting its slightly messy state indicated by creases and indentations, and contextualizes the scene within a residential bedroom with daylight filtering through partially open curtains.

- **Thoughts:** The model interprets the user's intent, identifying the dog's location post-interaction with the bed. Leveraging details from the object attributes summary (*e.g.,* the bed's usage state and decorative pattern) and spatial context from the visual description (residential setting, disturbed bed surface), the model infers that the dog's presence is closely related to the bed's recent disturbance and current state.

- **Reflections:** The model explicitly reflects on its reasoning steps, evaluating how the attributes "slightly messy" and patterned fabric provide critical visual and contextual anchors for retrieval. It also notes that the daylight and residential context reinforce the recent interaction scenario, logically concluding that the dog's probable location is directly on the bed itself.

- **Target Video Description:** The model synthesizes these insights into a concise and contextually coherent description: *"A dog standing on a slightly messy, patterned bed in a light-filled bedroom."*

This structured Reflective CoT approach enables the MLLM to systematically internalize reasoning patterns from textual exemplars, supporting consistent and accurate multimodal inference even without direct visual references. By utilizing language-only in-context demonstrations, our method effectively maintains training-free adaptability, enhancing retrieval accuracy through clearly articulated reasoning pathways.

A.6 CoT Pre-screening Prompt for Dataset Filtering

We screen all NLQ samples with a chain-of-thought prompt (As shown in the Figure 9) to keep queries whose main object is linguistically tied to the user (first-person possessives or deictics), while allowing secondary impersonal objects. For example, "What was the color of *my* drawstring bag?" is retained, whereas "In what aisle did I see a shopping trolley?" is excluded as a general, short-term lookup.

You are a helpful vision assistant that identifies object attributes in a structured JSON format.
Ensure your output is valid JSON with exactly the specified top-level keys.

Please carefully analyze the following sentence and identify the primary object being interacted with by the speaker.
The sentence is: "{object_sentence}"

Consider both the sentence and, if applicable, any associated visual information to determine the object and its attributes.

Your task is to describe this object's attributes in a structured JSON format.
The JSON output must contain EXACTLY the following top-level keys:
{required_keys_str} # Dynamically populated list, e.g., major_category, subcategory, color, texture, shape, material, brand, style, pattern, feature, usage, status.

Guidance for attribute values
1. Be Specific and Descriptive: For each attribute, provide the most accurate and detailed value you can infer. For example, for 'color', if an object is "dark blue with yellow stripes", please state that rather than just "blue" or "patterned".
2. Use Known Values as Examples: Below is a list of attribute categories and examples of values seen previously. Use these to understand the type of information expected. If a relevant value is present, you can use it.
Known attribute examples:
{known_str_joined} # Dynamically populated, e.g., "- color: (e.g., red, blue, green, ...)"
3. Invent New Values When Necessary: If the object has a characteristic not covered by the examples or if the examples are not relevant, provide a new, concise, and descriptive value. This is how we discover new attributes.
4. Handling Uncertainty/Not Applicable: If an attribute's value cannot be determined from the provided information or is not applicable to the object, use "unspecified" or "N/A". Avoid guessing if confidence is very low. For 'brand', if not explicitly mentioned or visible, "unspecified" is appropriate.
5. Compound Attributes: For attributes like 'material' (e.g., "fabric and wood") or 'feature' (e.g., "supporting and sleeping"), list the distinct components as a single string.
6. Consistency: Ensure your entire output is a single, valid JSON object with all string values properly escaped.

Example of desired thinking for 'color' if a bed is blue and white patterned:
"color": "blue and white patterned"

Now, based on the sentence "{object_sentence}" and any visual context, provide the JSON output.

Figure 6: Template used to construct the personalized memory bank.

You are a highly skilled AI assistant specializing in multimodal video retrieval with deep chain-of-thought reasoning. You will receive:

Visual Reference Image
– A single frame (the middle frame) from a candidate video.
– Contains key visual details: object color, shape, texture, size, and surrounding context (indoor/outdoor, lighting, background scene).

Object Attributes Summary
– A concise personal profile of the object, including categorical and frequency data (e.g., major category, subcategory, color combinations, material, style, usage frequency).
– Reflects the user's habitual usage and personal signature.

Your task is to produce a detailed chain-of-thought explanation and a final one-sentence description of the target video. The target video is assumed to contain the object referenced in the user's query, based on both the visual evidence from the candidate frame and the user's personal usage information.

Your response must be structured as a JSON object with the following keys:
{
  "Original Image Description": <string>,
  "Thoughts": <string>,
  "Reflections": <string>,
  "Target Video Description": <string>
}

Guidelines on Generating the Original Image Description
- Provide a thorough and detailed description of the visual reference image.
- Describe all visible elements in the reference image: the object's attributes (color, shape, texture, size), its immediate surroundings, and indoor/outdoor context.
- Be precise and comprehensive.

Guidelines on Generating the Thoughts
- Explain your understanding of the user's query and the object attributes summary.
- Detail which visual cues (e.g., dominant colors, materials, spatial relations) align with the personal profile
- Consider semantic aspects such as Location/Positioning, Object Attributes, Temporal Sequence, Presence/Absence and Action/Manipulation.
- Discuss which details in the candidate image were most influential in guiding your decision-making process.
- Conclude with how these insights were used to formulate your final target video description.

Guidelines on Generating the Reflections
- Summarize how the integration of the visual clues and the object attributes influenced your approach.
- Highlight the most influential details (e.g., material, environment) and why they confirm the candidate video's relevance.
- Explain how specific details (such as color, material, or setting) reinforced your decision.
- Concise meta-reasoning: justify key decisions that preserved coherence between the reference image, the attribute summary, and the retrieval goal. Highlight which visual or personal-usage cues were decisive.
- Reflect on the overall impact of these considerations in crafting a logically connected and visually coherent final description.

Guidelines on Generating the Target Video Description
- Provide a single, concise sentence that identifies the most likely video segment containing the referenced object.

Below is an example of the expected input and output formats:
…

Figure 7: Reflective CoT prompt template used in EgoRetriever.

Example Input:
<Input>
{
“User Query”: “Where was the dog after I laid the bed? ”
“Visual Reference”: [Attached image showing the middle frame from a video,],
“Object Attributes Summary”:
"subcategory: bed(48)
color: varied and patterned(11), varied(8), unspecified(7), blue(4), multicolor and patterned(4), unsure(4), blue and multicolor(3), multicolor(3), beige and green(1), blue and white(1), beige and patterned(1), white and blue(1)
shape: rectangular(33), rectangular and slightly padded(8), rectangular and cushioned(6), rectangular and slightly cushioned(1)
material: fabric and wood(32), wood and fabric(9), fabric and synthetic(4), wood(2), wood and synthetic(1)
texture: soft(11), soft and slightly patterned(10), smooth and slightly soft(7), soft and slightly wrinkled(6), soft and slightly smooth(3), smooth(3), soft and slightly textured(3), soft and rough(2), smooth and slightly worn(1), soft and smooth(1), quilted(1)
size: large(45), medium(3)
brand: unspecified(33), unsure(15)
style: household and resting(31), residential(15), residential and resting(1), household and sleeping(1)
pattern: decorative and patterned(19), solid(14), decorative(7), unspecified(4), varied(3), decorative and floral(1)
feature: supporting and sleeping(18), resting(14), comfortable and supporting(10), supporting and resting(4), comfortable and supportive(1), adjustable(1)
usage: resting and sleeping(19), sitting(12), resting and supporting(11), resting(2), interaction and placing(1), resting and lying(1), sleeping and resting(1), residential(1)
status: used(15), used and in use(14), used and slightly untidy(9), used and slightly messy(7), unspecified(1), used and slightly wrinkled(1), used and slightly organized(1)“
}

Example Output:
<Response>
{
"Original Image Description": "A large rectangular bed with a decorative patterned fabric cover rests in a light-filled bedroom. The bed's surface is slightly messy, showing creases and indentations, with wooden bedframe legs visible. The surrounding room has residential furnishings, including a nightstand and partially open curtains allowing daylight.",
"Thoughts": "The user's query focuses on the dog's location post-bed interaction. The Object Attributes Summary emphasizes the bed's 'used and slightly messy' status, 'decorative and patterned' design, and 'household and resting' context. The visual reference shows a residential bedroom setting with a disturbed bed surface, suggesting recent activity. The target video must encode both the bed's attributes (pattern, texture, usage state) and spatial context (bedroom) to align with retrieval model needs.",
"Reflections": "The 'slightly messy' status from attributes and visible creases in the image strongly indicates recent bed interaction. The patterned fabric (mentioned in both color/texture attributes and visual description) provides distinctive visual anchors for retrieval models. Daylight through curtains in the reference image confirms the 'light-filled' environment. These elements collectively suggest the dog remained on the bed after interaction, with the patterned texture and spatial context being critical retrieval cues.",
"Target Video Description": "A dog standing on a slightly messy, patterned bed in a light-filled bedroom."
}

Figure 8: Vision-by-language in-context learning examples for Reflective CoT.

You have samples of a query and the middle frame of the target video for an **egocentric video retrieval** task, where each query may or may not represent a personalization query.
Personalization queries involve questions about objects or details closely related to the user's personal belongings, personal actions, or personal spaces.
Analyze the given query **step-by-step**, thinking carefully about whether it reflects a personal connection or just a general interaction or observation.

Guidelines on Generating the **Thoughts**
1. **Identify** the object or event mentioned in the query.
2. **Consider** if the object/event typically belongs to, or is personally associated with, the user.
3. **Determine** if the query involves personal ownership, personal routine, or personal responsibility.
4. **Distinguish** between personal-related questions (about belongings, personal events, personal spaces) and non-personal general observation questions.

Output Instructions
Based on your Thoughts, output your answer in **JSON format**, clearly indicating `"YES"` if it is a personalization query, or `"NO"` if it is not, along with your reasoning.

Input Format
```json
<Input>
{
    "query": "<User_Query>",
    "target_video": "<Target_Video_Frame>"
}
```
Output Format
```json
<Response>
{
    "personalization": "YES" or "NO",
    "reason": "<Your detailed reasoning>"
}
```
Example 1
```json
<Input>
{
    "query": "What was the color of my drawstring bag?"
}
<Response>
{
    "personalization": "YES",
    "reason": "The query references a personal belonging ('my drawstring bag'), clearly indicating personal ownership."
}
```

Figure 9: CoT prompt used to pre-screen Ego4D-NLQ queries during benchmark construction.

Algorithm 1 Calculating Cross-User Object Attribute Similarity (Jaccard)

Let $\mathcal{U}_O = \{U_k \mid (U_k, A_k) \in \mathcal{D} \text{ for object } O\}$ be the set of all users with attributes for object O.
Initialize a list of similarity scores $S = []$.
for each unique pair of users (U_i, U_j) in $\mathcal{U}_O$ where $i \neq j$ **do**
    Retrieve attribute sets for each user: $A_i = \mathcal{D}(U_i, O)$ and $A_j = \mathcal{D}(U_j, O)$.
    Compute the Jaccard similarity:

$$s_{ij} = J(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

    Add s_{ij} to S.
end for
if S is not empty **then**
    Compute the average similarity: $\text{Sim}_{\text{avg}}(O) = \frac{1}{|S|} \sum_{s \in S} s$
    Compute the average diversity: $\text{Div}_{\text{avg}}(O) = 1 - \text{Sim}_{\text{avg}}(O)$
else
    Set average diversity to 0 (or undefined if only one user has the object): $\text{Div}_{\text{avg}}(O) = 0$
end if
return $\text{Div}_{\text{avg}}(O)$

Algorithm 2 Computing Process of EgoRetriever for Personal Egocentric Video Retrieval

Input: Natural language query Q, candidate clip set $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$, user-specific semantic memory bank $\mathcal{M}$, reflective CoT prompt p_c.
Parameters: Frozen Multimodal Large Language Model (MLLM) Ψ_M, frozen text encoder Ψ_T, frozen video encoder Ψ_V.
Output: Retrieved target clip C^*.

Initialize pre-trained and frozen models Ψ_M, Ψ_T, Ψ_V.
Retrieve personal context from memory bank: Query $\mathcal{M}$ with Q to obtain personal object metadata $\mathcal{M}_q$ and reference image I_r.
Generate target clip description using reflective CoT:

$$T_t = \Psi_M(p_c \circ Q \circ \mathcal{M}_q \circ I_r)$$

Compute normalized text embedding for the target description:

$$\hat{e}_T = \frac{\Psi_T(T_t)}{\|\Psi_T(T_t)\|_2}$$

Initialize a list of similarity scores $S = []$.
for each candidate clip C_i in $\mathcal{C}$ **do**
    Compute normalized video embedding for the candidate clip:

$$\hat{e}_{V_i} = \frac{\Psi_V(C_i)}{\|\Psi_V(C_i)\|_2}$$

    Compute similarity score: $s_i = \hat{e}_{V_i}^\top \hat{e}_T$ {Cosine similarity for normalized embeddings}
    Add s_i to S.
end for
Retrieve target clip: $C^* = \underset{C_i \in \mathcal{C}}{\text{argmax}}\, S_i$ {Select clip with highest similarity score}
return C^*

B EgoMemory Benchmark Details and Statistics

B.1 Benchmark Construction and Filtering

To rigorously evaluate memory-augmented *personalized* egocentric video retrieval, we introduce the **EgoMemory** benchmark, constructed from Ego4D's Natural Language Queries (NLQ) (Grauman et al., 2022). NLQ provides ∼227 hours of head-mounted video from 137 participants across 74 locations, with free-form queries about "when/where/with whom/what," closely reflecting realistic recall scenarios.

While NLQ offers rich content, its original design targets temporal localization within individual clips and does not aggregate multi-video, user-specific context. We therefore make personalization *explicit*. Concretely, we operationalize personalization as *personally experienced objects*: entities that (i) recur across a user's videos and/or (ii) are linguistically tied to the user via first-person/possessive/deictic cues. This definition does not require legal ownership; instead, it uses recurrence and linguistic evidence to avoid conflating incidental scene exposure with genuine user linkage.

To address NLQ's gaps, our benchmark treats each user as a distinct retrieval unit, aggregating all of their videos as long-term context at query time. We apply a three-stage filtering pipeline: (1) **GPT-4o CoT pre-screening** to retain queries with explicit personal references as shown in Figure 1; (2) **long-context participant selection**, requiring at least 10 videos and ≥1 hour cumulative footage per user to ensure sufficient temporal breadth; (3) **manual verification**, where annotators review 20 additional clips (3s each) from the same user for the queried object class and label a query "personal" if ≥90% of reviewed instances match ("uncertain" if ≥75%). This yields 639 curated queries over 45 participants (245 videos), with ∼91.6% labeled personal, thus emphasizing queries whose resolution benefits from cross-video personal cues rather than single-clip idiosyncrasies. Specifically:

1. **GPT-4o CoT pre-screening.** We screen all NLQ samples with a chain-of-thought prompt (As shown in the Figure 9) to keep queries whose main object is linguistically tied to the user (first-person possessives or deictics), while allowing secondary impersonal objects. For example, "What was the color of *my* drawstring bag?" is retained, whereas "In what aisle did I see a shopping trolley?" is excluded as a general, short-term lookup.

2. **Long context participant selection.** We define long context as users having at least 10 videos and ≥1 hour cumulative duration. For retained queries, the referenced main object must recur across a user's videos, ensuring that answering the query draws on cross-video history rather than a single clip.

3. **Manual verification.** For each query–target pair, annotators inspect 20 additional 3 s clips from the same user containing the same object class. A query is "personal" if ≥90% of reviewed instances match the target object and "uncertain" if ≥75%. This process results in ∼91.6% personal queries and reduces conflation with scene exposure.

To address these gaps, our benchmark treats each user as a distinct retrieval unit, aggregating all of their available videos as personal context at query time. We apply rigorous data filtering: participants must have multiple videos, and we use GPT-4o to select queries with long-context dependencies, followed by manual curation (see Supplementary for details). The resulting dataset comprises 245 videos from 45 unique participants, spanning diverse everyday contexts and totaling 639 distinct queries.

Personalized memory banks are constructed for each participant as described in Section 3.2.1, resulting in 165,795 user-specific object annotations. The number of unique object types per participant ranges from 59 to 638 (median: 129), with memory bank sizes ranging from 322 to 10,454 annotations (median: 1,312). Figure 12 visualizes the most frequent objects.

The candidate retrieval set includes 2,228 video clips extracted from participants' historical egocentric videos. Each user's average candidate set size is approximately 33 clips, capturing a rich spectrum of personal and habitual contexts. Video lengths within EgoMemory span from as short as 4 seconds to a maximum of 300 seconds, averaging 103.82 seconds per clip. The total cumulative duration of the candidate set amounts to 64.25 hours. Notably, these varied temporal spans enable thorough assessments of retrieval robustness across different episodic memory lengths and complexities.

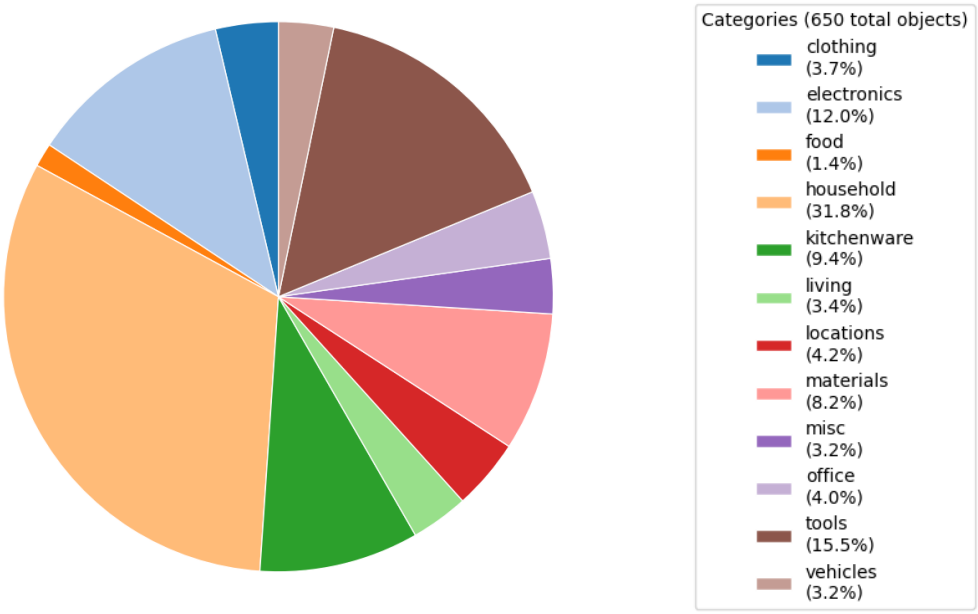

Figure 10: Object-category distribution in the candidate retrieval pool.

B.2 Candidate Pool Statistics

To facilitate personalized episodic memory retrieval, we construct a candidate retrieval set for each participant by aggregating their historical egocentric video clips prior to each query. This design simulates realistic personal memory banks and enables a thorough evaluation of retrieval models across diverse user contexts.

To facilitate personalized episodic memory retrieval, we construct a candidate retrieval set for each participant by aggregating their historical egocentric video clips prior to each query. This design simulates realistic personal memory banks and enables a thorough evaluation of retrieval models across diverse user contexts.

Composition and Scale. The benchmark candidate set comprises a total of 2,228 video clips collected from 45 unique users, sampled from their continuous head-mounted video recordings. The number of candidate clips per participant ranges from 9 to 61, with an average of 33 clips per user. Clip durations vary from 4 to 300 seconds, with a mean duration of 103.82 seconds, resulting in a cumulative total of 64.25 hours of video data in the candidate pool.

Object Instance Distribution. Within the candidate set, we annotated 650 object instances spanning 13 semantic categories. Figure 10 illustrates the distribution of these object categories. The largest proportions are household objects (31.8%), tools (15.5%), and electronics (12.0%), alongside kitchenware (9.4%), materials (8.2%), and other everyday object classes. This distribution reflects the diversity and complexity of real-world personal environments encountered in egocentric video.

Personalization and Diversity. To assess diversity, we computed Jaccard similarity scores for the top 100 most frequent object types based on their attribute metadata. Inter-participant similarity was below 0.4 for 74% of object types, indicating high heterogeneity between users. Furthermore, intra-participant similarity was consistently higher than inter-participant similarity, underscoring the personalized and user-specific nature of each candidate set.

Evaluation Metric. For all experiments, we report **mean Recall@K** (K=1, 2, 3), computed separately for each user over their candidate set and then macro-averaged across users. This approach, consistent with prior benchmarks (Grauman et al., 2022; Hummel et al., 2024), ensures that performance reflects the retrieval difficulty for each individual and mitigates biases due to uneven query counts. In cases with a single correct answer per query, Recall@K is equivalent to Hit Rate@K (Sun et al., 2019), a standard metric in recommender systems.

B.3 Memory-bank Statistics and Diversity

B.3.1 Object Distribution Visualization

To provide a more intuitive understanding of the environment and objects present in our EgoMemo-

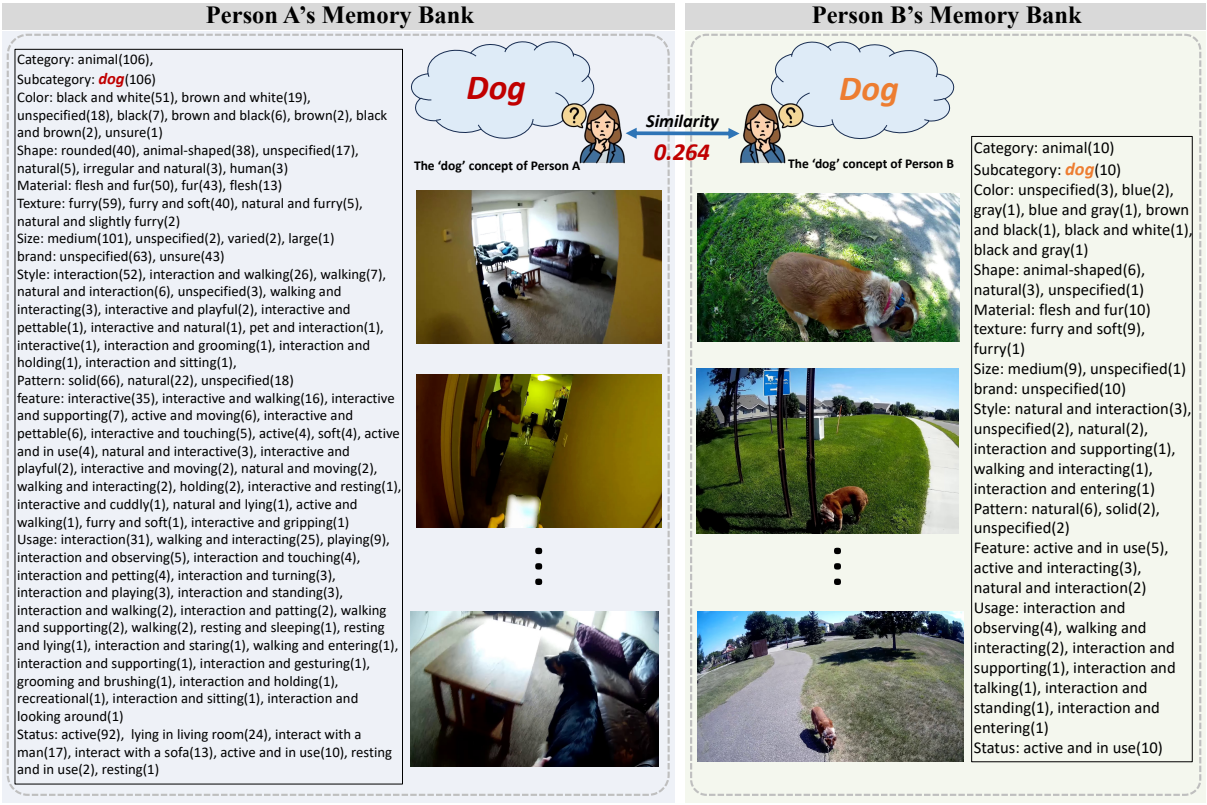

Figure 11: Qualitative comparison of the object concept "dog" across two users' personal memory banks. Attribute distributions reveal substantial divergence in visual, contextual, and interactional properties. The computed similarity score (Jaccard index) is **0.264**, highlighting the necessity of modeling user-specific memory to capture personalized object semantics.

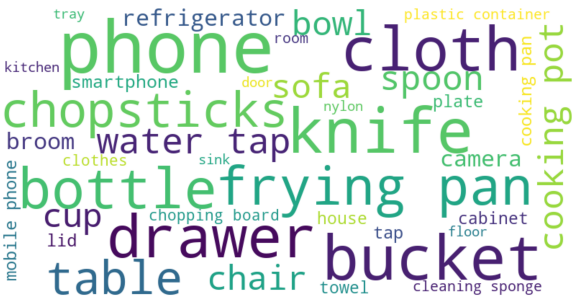

Figure 12: Most frequent objects across memory banks in EgoMemory.

rybenchmark, we present a word cloud visualization of the most frequently occurring objects in Figure 12.

The visualization highlights that the benchmark is grounded in daily activities, featuring common household and personal items such as *phone*, *knife*, *table*, and *drawer*. This diversity ensures that the personalized retrieval task covers a wide range of realistic interaction scenarios.

B.3.2 Qualitative Analysis for Personalized Memory Bank

To better understand the practical implications of our proposed personalized memory bank, we conducted a qualitative analysis comparing attribute representations of identical object categories across different users. As an illustrative example in Figure 11, we examined the concept of "dog" as represented by two distinct users (Person A and Person B), each with their own historical interactions captured within their respective personal memory banks.

The detailed attribute annotations reveal substantial variations between the two users in aspects such as color, pattern, style, usage, and specific interaction contexts. Quantitatively, the computed similarity score between these two users' personal memory for the concept of "dog" (Indoor *vs* Outdoor) is notably low, at **0.264**. This underscores a significant divergence in their individual conceptualizations and experiences associated with the same general object class.

The low similarity score highlights a crucial in-

Method	Video Model	Textual Memory Bank	Visual Reference	Fusion Strategy	Mean Recall (%) mR@1	mR@2	mR@3
Random	✗	✗	✗	—	3.62	9.74	15.23
CLIP	✗	✓	✗	—	10.41	12.95	16.72
BLIP	✗	✓	✗	—	10.88	13.67	17.48
EgoVLPv2	✓	✓	✗	—	11.25	15.03	18.30
LanguageBind	✓	✓	✗	—	11.02	14.60	17.83
CLIP	✗	✗	✓	Avg	14.74	17.53	21.41
BLIP	✗	✗	✓	Avg	15.12	18.07	22.82
EgoVLPv2	✓	✗	✓	Avg	15.77	20.61	23.79
LanguageBind	✓	✗	✓	Avg	15.26	20.14	23.04
CLIP	✗	✓	✓	Avg	15.64	18.63	22.71
BLIP	✗	✓	✓	Avg	16.02	19.17	24.12
EgoVLPv2	✓	✓	✓	Avg	16.67	21.71	25.09
LanguageBind	✓	✓	✓	Avg	16.16	21.24	24.34
$\text{BLIP}_{\text{CoVR}}$ (Ventura et al., 2024)	✗	✓	✓	Cross-Attn.	15.94	19.17	23.00
$\text{BLIP}_{\text{CoVR-ECDE}}$ (Thawakar et al., 2024)	✗	✓	✓	Cross-Attn.	16.41	19.63	23.64
CIReVL (Karthik et al., 2024)	✗	✓	✓	Captioning	16.95	20.13	24.37
OSrCIR (Tang et al., 2025b)	✗	✓	✓	Captioning	17.28	21.64	25.49
VideoAgent (Fan et al., 2024)	✓	✓	✓	Captioning	17.49	26.40	35.62
TFR-CVR (Hummel et al., 2024)	✓	✓	✓	Captioning	18.21	27.12	32.05
VideoRAG (Luo et al., 2025)	✓	✓	✓	Captioning	<u>19.70</u>	<u>30.83</u>	<u>39.82</u>
EgoRetriever (Ours)	✓	✓	✓	Captioning	**23.19**	**38.48**	**47.83**

Table 7: Detailed mean Recall@K (%) for different retrieval configurations. This table supplements the main results by detailing the impact of textual and visual inputs.

sight: object attributes are perceived and recalled uniquely by different users based on their personal experiences and contexts. Thus, it clearly demonstrates the necessity and importance of constructing personalized memory banks, as generic or aggregated memory representations would inadequately capture the rich variability in individual user interactions and perceptions. Our findings reinforce the core contribution of our personalized memory bank framework, its ability to accurately reflect nuanced user-specific memory contexts, ultimately enhancing personalized retrieval performance.

C Additional Experimental Results on EgoMemory

C.1 Full Results Across Input Modalities

Table 7 presents the complete performance breakdown across different input regimes: (a) query plus textual memory, (b) query plus visual reference, and (c) fusion. The results demonstrate that models relying solely on textual memory are limited in capturing user intent (*e.g.,* EgoVLPv2 mR@1: 11.25%), highlighting the necessity of integrating visual references. The combination of textual memory and visual reference leads to consistent gains across all baselines.

C.2 Memory-bank Design Ablations

Table 8 reports retrieval performance under different memory bank structures. Across both TFR-CVR and EgoRetriever, our structured metadata yields clear gains over caption memories from GPT–4o and EgoGPT. For TFR-CVR, metadata improves average performance by 9.79% and 7.87% relative to GPT-4o and EgoGPT, respectively. For EgoRetriever, the gains are larger at 18.62% and 15.63%. We also evaluate the VideoAgent memory bank (Fan et al., 2024), which uses a temporal caption memory and an object memory, and find it trails our structured metadata and is comparable to EgoGPT for TFR-CVR. We attribute the advantage to storing 12 attribute fields and frequency statistics that emphasize recurrent personal objects, enabling richer long-context grounding than caption-only or non-personalized memories.

C.3 Additional Ablations

Table 9 presents further ablation results that complement the analysis in the main paper by evaluating the significance of our one-stage reasoning pipeline, the role of different CoT variants, and reference frame selection strategies.

(1) Models '2–3' assess the significance of the one-stage reasoning strategy. We compare our unified reflective inference process with a two-

Method	Memory Structure	mR@1	mR@2	mR@3
TFR-CVR	GPT-4o Caption	11.41	16.57	20.04
	EgoGPT Caption	13.09	19.37	21.32
	VideoAgent	13.09	19.37	21.32
	Metadata	18.21	27.12	32.05
EgoRetriever	GPT-4o Caption	13.18	19.30	21.17
	EgoGPT Caption	15.31	20.57	26.74
	VideoAgent	17.49	26.40	35.62
	Metadata	**23.19**	**38.48**	**47.83**

Table 8: Results emphasizing the importance of our personal memory bank structure.

stage pipeline where the reference caption and final description are generated sequentially. The two-stage baseline (model '2') yields an mR3 of 32.05, which is 15.78% lower than the full one-stage model (model '1'). Incorporating basic Chain-of-Thought prompting into the two-stage pipeline (model '3') improves performance to 40.73, yet still underperforms our one-stage reflective reasoning, indicating the effectiveness of jointly reasoning over query, reference, and memory in a single coherent pass.**(2) Models '4–6' explore alternative prompting strategies for reflective reasoning.** Replacing our Reflective CoT with a simple, flat instruction (model '4') or a simple CoT strategy (model '5') leads to 6.21% and 4.97% average performance drops, respectively, confirming the necessity of deep multimodal reasoning. Further, we evaluate a structured CoT variant using DDCoT (Zheng et al., 2023), which first decomposes user queries before composing descriptions (model '6'), resulting in the lowest performance among CoT variants. This suggests that sequential decomposition may hinder holistic interpretation in personalized video contexts, reinforcing the advantage of our proposed Reflective CoT.**(3) Models '7–8' evaluate alternative reference frame selection strategies.** Compared to the randomly selected frame baseline (model '7'), our middle-frame selection strategy (model '1') provides a slightly improved retrieval performance (0.427% higher on average). An alternative frame selection method, which leverages clip embeddings to calculate the similarity between captions and individual frames (model '8'), achieves marginally higher performance (0.313% higher on average). However, this embedding-based approach significantly reduces efficiency due to the computational overhead of evaluating similarity against dense frame-level descriptions provided by Ego4D's narrations (at 30 fps). Therefore, our simpler middle-frame strategy is optimal, striking an effective balance between retrieval accuracy and computational efficiency. **(4) Model '9-10' analysis the dependence on human narrations of our personal memory bank.** Dropping 25% (model '9') and 50% (model '10') of human narrations yields moderate degradations, consistent with redundancy in our memory bank: recurring personal objects are recorded across clips and frequency-based filtering suppresses noise from missing narrations.

C.4 Additional Qualitative Examples

To demonstrate the advantages of reflective CoT in accurately interpreting user intent, we present qualitative analyses in Figure 13. Reflective CoT plays a crucial role in filtering out irrelevant elements, such as extraneous scene descriptions or hallucinated details, which are often distractions in the reasoning process. This reflective process enhances the precision and reliability of episodic memory retrieval. In Row 1, reflective CoT excels in focusing on the relevant context by identifying the state and placement of gloves before being dropped on a table. Despite potential visual clutter, reflective CoT correctly infers the gloves' usage context, i.e., being worn and involved in manual tasks, reinforcing the memory of the gloves as protective gear. Row 4 further illustrates the power of reflective CoT by eliminating an incorrect assumption in the reasoning process. The image initially suggests indoor parking and a black bicycle, but reflective CoT shifts focus to correctly interpret the user's query about interactions during active bicycle usage outdoors. This change in focus underscores the adaptability of reflective CoT in aligning with the user's true intent, effectively disregarding irrelevant contextual details.

By integrating personal memory and contextual cues through explicit reflection, reflective CoT ensures a robust understanding of the user's intent. This not only filters out extraneous elements but also contributes to more accurate retrieval of relevant episodic memories, ultimately enhancing the system's interpretability and performance.

D Generalization Benchmarks

D.1 EgoCVR: Setup and Results

D.1.1 Overview

To further assess the generalization ability of our proposed EgoRetriever framework, we conduct additional experiments on the EgoCVR bench-

	Methods	mR@1	mR@2	mR@3
1.	Full Model (Reflective CoT, one-stage, middle reference frame)	23.19	38.48	47.83
Significance of the one-stage reasoning strategy				
2.	Two-stage	18.21	27.12	32.05
3.	Two-stage + CoT	19.92	32.41	40.73
Alternative prompting strategies for Reflective CoT				
4.	Simple prompt	19.27	31.48	40.11
5.	Simple CoT (no reflection and designed thoughts)	20.13	33.18	41.28
6.	DDCoT (Zheng et al., 2023)	19.42	32.50	41.02
Alternative reference frame select strategies for Reflective CoT				
7.	Random selected	23.04	38.02	47.16
8.	Context-based selected	23.39	38.94	48.11
Dependence on human narrations				
9.	25% narrations dropped	22.46	36.21	46.09
10.	50% narrations dropped	21.82	35.19	45.32

Table 9: Additional ablation results in terms of mR@1, mR@2, and mR@3, demonstrating the necessity of one-stage reasoning and the superiority of our Reflective CoT design.

mark (Hummel et al., 2024). EgoCVR is an egocentric, composed video retrieval dataset designed to evaluate a model's capability to generate target video descriptions conditioned on reference images and textual modifications. This benchmark presents unique challenges by emphasizing nuanced temporal and semantic variations in first-person video data, complementing our primary evaluations on EgoMemory.

D.1.2 Task Definition.

In EgoCVR, the Composed Video Retrieval (CVR) task is defined as follows: Given a query comprising a reference frame (sampled from a video) and a free-form textual modification instruction, the objective is to retrieve the corresponding target video clip from a gallery. This retrieval is performed under two conditions: a *global* setting, where the gallery includes all test clips, and a *local* setting, where the gallery is restricted to clips from the same long-form video as the query. Formally, let $\mathcal{I}$ denote the set of reference images, $\mathcal{T}$ the set of textual modifications, and $\mathcal{V}$ the set of candidate videos. For each query (i_q, t_q), the task is to identify $v^* \in \mathcal{V}$ that best reflects the semantic transformation specified by t_q when applied to i_q.

D.1.3 Dataset Construction.

The EgoCVR dataset consists of egocentric videos capturing a diverse array of daily activities and object manipulations. Each annotated instance comprises:

- A **reference image**: a single frame extracted from the query video.
- A **textual modification**: a concise natural language instruction describing the intended change (*e.g.*, "use a different object," "perform the action faster").
- A **target video**: a short clip (2–8 seconds) from the dataset that realizes the specified modification.

Each query is paired with ground-truth target(s), and standard test splits are provided to ensure comparability across methods. For further details on dataset construction, we refer the reader to (Hummel et al., 2024).

D.1.4 Evaluation Protocol

We follow the official evaluation protocol proposed in (Hummel et al., 2024). Specifically, two retrieval settings are considered:

- **Global Setting:** The retrieval gallery consists of all test video clips, simulating a large-scale retrieval scenario with numerous visually and semantically similar distractors.
- **Local Setting:** The retrieval gallery is restricted to clips extracted from the same long video as the reference, enabling evaluation of fine-grained, within-context retrieval.

For each query, the model generates a target description based on the reference image and textual modification, then retrieves the most relevant clip from the gallery.

D.1.5 Evaluation Metrics

Performance is measured using standard top-k retrieval metrics:

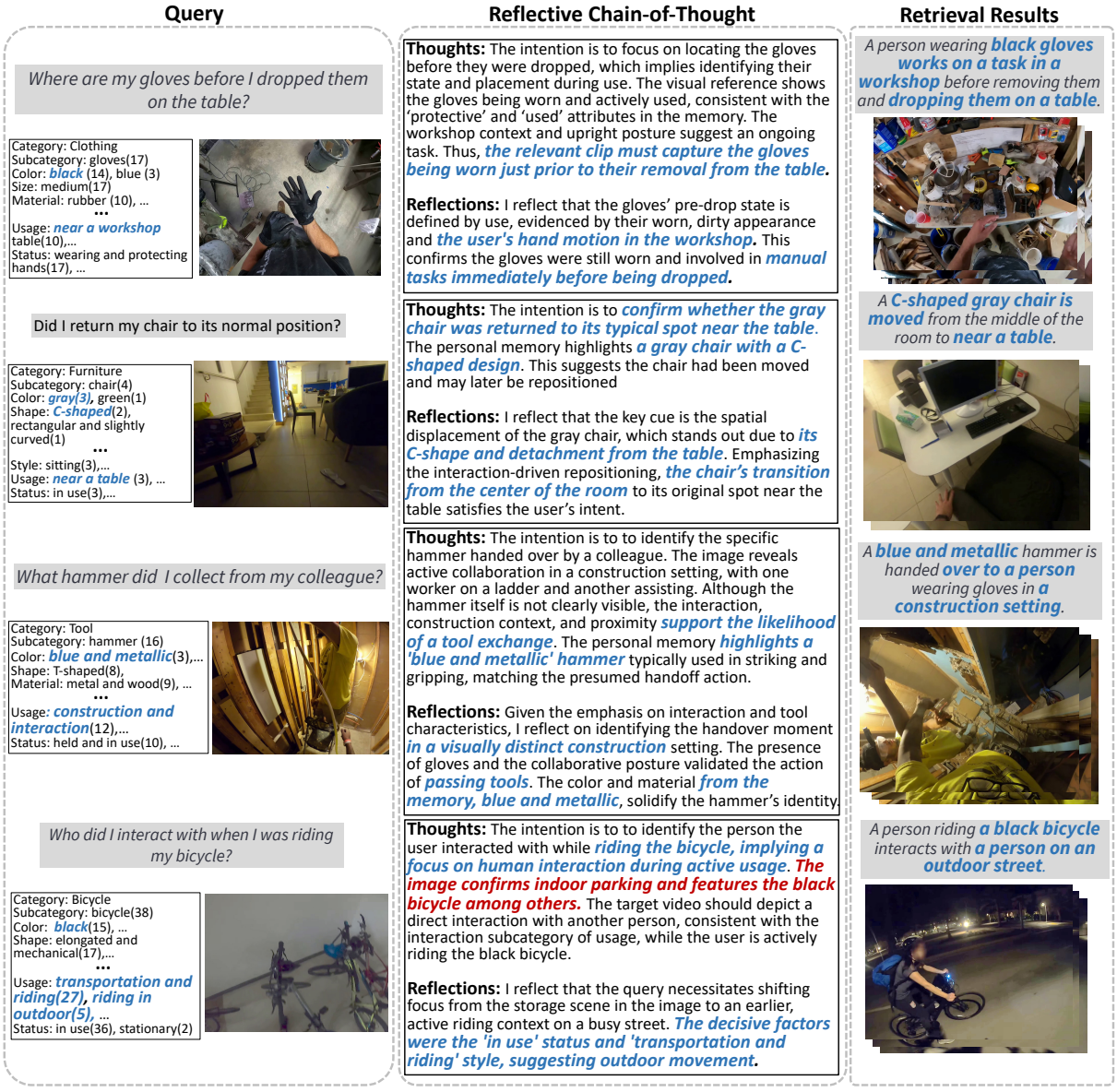

Figure 13: Additional qualitative examples of Reflective CoT on EgoMemory.

- **Recall@K (R@K):** The proportion of queries for which at least one ground-truth target appears among the top K retrieved results. Higher values indicate better retrieval performance.

Following the official setting (Hummel et al., 2024), we report R@1, R@5, and R@10 for the global setting, and R@1, R@2, and R@3 for the local setting, following the original benchmark protocol. Metrics are averaged across all test queries.

D.1.6 Experimental Setup

All experiments are conducted on the official EgoCVR test splits, strictly adhering to the standardized evaluation settings. Model implementations and hyperparameters follow those described in Section 4 of the main paper. The primary methods compared include:

- **CIReVL** (Karthik et al., 2024): A training-free composed retrieval approach adapted from image retrieval to video.
- **OSrCIR** (Tang et al., 2025b): A one-stage, training-free composed retrieval framework.
- **TFR-CVR** (Hummel et al., 2024): A two-stage approach using video captioning followed by LLM-based query modification.
- **TFR-CVR***: A variant of TFR-CVR using GPT-4o as both captioner and modifier.
- **VideoAgent** (Fan et al., 2024): An agentic

Table 4: Analysis the generalization of our EgoRetriever on the EgoCVR benchmark.

Method	Global			Local		
	R@1	R@5	R@10	R@1	R@2	R@3
CIReVL (Karthik et al., 2024)	2.0	6.8	10.6	33.6	49.7	61.4
OSrCIR (Tang et al., 2025b)	4.9	9.3	13.4	37.4	53.3	68.1
TFR-CVR (Hummel et al., 2024)	14.1	39.5	54.4	44.2	61.0	73.2
TFR-CVR*	14.7	41.2	55.6	46.1	62.4	73.9
EgoRetriever	**17.4**	**49.2**	**62.7**	**50.3**	**68.2**	**76.4**

long-video understanding framework that decomposes the query into sub-tasks and iteratively retrieves and verifies supporting evidence from video before producing the final answer.

- **VideoRAG** (Luo et al., 2025): A retrieval-augmented long-video comprehension method that performs visually-aligned evidence retrieval and conditions generation on the retrieved video segments.

- **EgoRetriever**: Our proposed one-stage, training-free framework that leverages MLLMs and reflective chain-of-thought reasoning to generate precise target descriptions for retrieval.

D.1.7 Results

As summarized in Table 4 of the main paper, EgoRetriever achieves state-of-the-art performance on EgoCVR across both global and local retrieval settings. Our method consistently outperforms all baselines in Recall@K, highlighting its superior ability to generalize to novel egocentric video retrieval tasks.

D.1.8 Comparison of Input Modalities: Our Benchmark vs. EgoCVR

While both our proposed benchmark (EgoMemory) and the EgoCVR benchmark (Hummel et al., 2024) focus on the challenging task of composed video retrieval, a key distinction lies in the design of their input modalities and the manner in which user intent and contextual information are encoded for retrieval.

Our Benchmark (EgoMemory): Our evaluation protocol is motivated by real-world episodic memory retrieval, wherein the user's search intent is inherently personal and context-dependent. To capture this, we design queries to include three components:

- **Textual Memory Bank:** A collection of personal, user-centric text snippets that serve as a long-term memory repository, reflecting frequently encountered objects, past actions, or unique user experiences. This memory bank enables models to retrieve and reason over personalized historical context during retrieval.

- **Reference Image:** A visual snapshot or frame representing the starting point of the query, grounding the search in a specific visual context.

- **User Query:** A free-form natural language request, which may include references to personal context, intent, or temporal cues (*e.g.,* "Find when I last used the red mug in the kitchen").

This multimodal input setting allows models to perform retrieval that is both visually grounded and deeply personalized, supporting rich reasoning over long-context egocentric video archives.

EgoCVR Benchmark. By contrast, the EgoCVR benchmark (Hummel et al., 2024) employs a more constrained input format, where each query consists of:

- **Reference Image:** A single frame from the query video, serving as the visual anchor for retrieval.

- **Modification Text:** A concise natural language instruction specifying a semantic change or action to be performed (*e.g.,* "pick up a different object," "use the other hand").

Notably, EgoCVR does not provide explicit access to personalized memory or long-term user context; all retrievals are based solely on the visual and textual modifications present in the immediate query. This design tests the model's ability to interpret and execute fine-grained semantic transformations but does not directly address personalization or long-horizon reasoning.

Summary of Differences. The primary distinction is that EgoMemory introduces a *textual memory bank* and *user query* to explicitly model personal context and long-term memory, supporting retrieval scenarios that are more representative of real-world egocentric memory augmentation. In contrast, EgoCVR focuses on visually-anchored

modifications without leveraging historical or personalized information. Our experimental evaluation on both benchmarks demonstrates the generalization of our approach to settings with and without access to external memory, highlighting the flexibility and robustness of our retrieval framework.

D.2 EgoLifeQA: Setup and Results

D.2.1 Overview

EgoLife is a week-long, multi-person egocentric study in which six participants cohabit and continuously record with AI glasses, producing a comprehensive ∼300-hour multimodal dataset; EgoLifeQA is a derived QA suite tailored to long-context assistance. The released EgoLifeQA subset contains 6,000 questions over 266 hours of video.

D.2.2 Task Definition

EgoLifeQA defines five question types: *EntityLog* (objects and their attributes/locations), *EventRecall* (past event details), *HabitInsight* (personal habit patterns), *RelationMap* (interpersonal interactions), and *TaskMaster* (task tracking and reminders). Each item is multiple-choice and is constructed to require evidence from at least five minutes prior to the question time.

D.2.3 Dataset Construction

Long, "visual–audio" captions are first generated and fed to GPT-4o to propose timestamped QA candidates; annotators then filter and refine them, retaining only questions with long-range dependencies (at least five minutes) and high real-world relevance, yielding the final QA set.

D.2.4 Evaluation Protocol

We follow the official setting: models must answer the fixed-choice questions using only the benchmark inputs and produce supporting predictions per category.

D.2.5 Evaluation Metrics

Performance is reported as accuracy (%) per question type, averaged over all five official categories (EntityLog, EventRecall, HabitInsight, RelationMap, TaskMaster).

D.2.6 Experimental Setup

Implementations and hyperparameters mirror those in the main paper. We compare EgoRetriever to **GPT-4o**, **LLaVA-OV**, and **EgoGPT** under identical inputs; EgoRetriever uses the same MLLM backbone (GPT-4o) and our reflective reasoning to form answers.

D.2.7 Results

As summarized in Table 2, EgoRetriever attains the highest accuracy on all five EgoLifeQA categories: **EntityLog** 42.5, **EventRecall** 45.1, **HabitInsight** 37.7, **RelationMap** 35.8, and **TaskMaster** 46.2, outperforming the strongest baseline (EgoGPT) by +3.3, +8.6, +6.6, +2.2, and +6.5 points, respectively, with a macro-mean of 41.5 vs. 36.0 (+5.5). These gains demonstrate transfer beyond retrieval to fixed-answer, long-context QA spanning entities, events, habits, interpersonal relations, and task tracking.

E Efficiency, Cost, and Scalability

E.1 Cost and Latency Breakdown

In Table 5 we report a four-way cost–effectiveness analysis, mean query latency (averaged over 100 runs), peak GPU memory at inference, OpenAI API expenditure per call, and average retrieval accuracy (ViT-L/14 backbone) on EgoMemory and EgoCVR, covering two prior baselines (CIReVL and OSrCIR), the two-stage TFR-CVR, and our single-stage EgoRetriever variants. CIReVL, which relies on a GPT-3.5 captioning head plus a separate retrieval LLM, incurs the highest memory footprint (40 GB) and the slowest response time (∼1.4 s), despite its modest API fee, and delivers the lowest accuracy (24.86). OSrCIR and TFR-CVR both reduce memory to 16 GB by avoiding a BLIP-2 captioner; however, OSrCIR's one-stage design halves latency to ∼0.7 s but still trails TFR-CVR by 13.4 mR points. Our EgoRetriever with GPT-4o-mini preserves OSrCIR's low memory/latency profile while lifting performance to 46.92, and the full GPT-4o variant maintains this latency ($\sim 0.7 \pm 0.08s$) while achieving the highest accuracy (48.19) without increasing memory consumption. Taken together, these results show that (i) eliminating separate captioning modules yields substantial memory and speed benefits, (ii) mini-scale MLLMs already strike an excellent cost–performance balance, and (iii) our reflective one-stage pipeline offers the best absolute accuracy with competitive operational costs, underscoring its suitability for real-time, resource-constrained egocentric retrieval systems.

Model	LLM	Latency	GPU Memory	API Cost	Avg. Performance
CIReVL	GPT-3.5	~ 1.4s	40 GB	$\sim$ \$0.001	24.86
OSrCIR	GPT-4o	$\sim 0.7 \pm 0.08$s	16 GB	$\sim$ \$0.004	27.87
TFR-CVR	GPT-4o	~ 1s	16 GB	$\sim$ \$0.007	41.25
EgoRetriever	GPT-4o-mini	$\sim 0.5 \pm 0.05$s	16 GB	$\sim$ \$0.002	46.92
EgoRetriever	GPT-4o	$\sim 0.7 \pm 0.08$s	16 GB	$\sim$ \$0.004	48.19

Table 5: Comparative analysis of computational cost, latency, memory usage, API expenditure, and retrieval performance across baseline and proposed models.

# Videos	Avg. Latency (s)	GPU Memory (GB)	API Cost (USD)
10	0.65 ± 0.03	16	$\sim$ \$0.004
50	0.70 ± 0.05	16	$\sim$ \$0.004
100	0.70 ± 0.07	16	$\sim$ \$0.004
245	0.70 ± 0.08	16	$\sim$ \$0.004

Table 6: Scalability with dataset size. Latency is averaged over 100 queries; API cost is per call.

E.2 Scalability with Increasing Video/Metadata Volume

As Table 6 shows, query-time latency remains $\approx$0.65–0.70 s with small variance as the corpus grows from 10 to 245 videos; peak GPU memory is flat at 16 GB and per-query API expenditure is unchanged. This stability stems from EgoRetriever using precomputed video and memory embeddings, so online retrieval reduces to lightweight lookups and similarity scoring; only offline indexing scales with data volume and does not affect interactive latency. In tandem with Fig. 6 (main), where larger memory-bank context improves accuracy, these results indicate that EgoRetriever scales efficiently in both cost and effectiveness as personalized data accumulates.

F More Implementation Details, Extended Related Work, and Broader Impacts

F.1 More Implementation Details

VideoAgent. We reproduce VideoAgent (Fan et al., 2024) under our captioning-based retrieval protocol. Following its design, we construct a two-part memory bank for each user: (i) a *temporal caption memory* that stores an MLLM-generated caption for each short video segment (2s granularity), and (ii) an *object memory* that tracks salient objects/people with their occurrences (and coarse locations when available) across the history. At inference time, the agent iteratively queries these memories to gather evidence relevant to (Q, I_r), refines its hypothesis, and outputs a single target description T_t intended to describe the target clip. We then rank all candidates by cosine similarity between T_t and candidate clip text/video embeddings, using the same pretrained video-language encoder as other captioning baselines (*i.e.,* EgoVLPv2) for fairness. We cap the agent interaction to a fixed budget (maximum S reasoning rounds and top-K memory entries per round) to control latency.

VideoRAG. We implement VideoRAG (Luo et al., 2025) as a retrieval-augmented generation baseline adapted to output a retrieval query in the form of a target description. Specifically, we index each candidate clip with its MLLM-generated caption (and optionally lightweight object tags), forming an evidence corpus. Given (Q, I_r), VideoRAG first retrieves the top-K evidence clips/segments by dense similarity in the caption embedding space. An MLLM (*i.e.,* GPT-4o) is then conditioned on the retrieved evidence together with (Q, I_r) to synthesize a target description T_t that consolidates the most relevant constraints. Finally, we rank the full candidate set using the same cosine-similarity retrieval backend as other captioning baselines, ensuring that any gains stem from the RAG-style evidence selection and aggregation rather than changes to the retrieval encoder.

Rigorous Data Filtering To ensure rigorous evaluation and the authenticity of personalized retrieval contexts in the **EgoMemory** benchmark, we implemented a comprehensive data filtering process. Initially, we screened the Ego4D Natural Language Queries (NLQ) dataset (Grauman et al., 2022), restricting inclusion criteria to participants possess-

ing multiple egocentric videos. This criterion was essential to authentically simulate realistic retrieval scenarios, as genuine personal retrieval contexts inherently involve multiple video interactions over extended periods.

Next, we utilized GPT-4o to systematically identify and select queries exhibiting clear long-context dependencies. Specifically, GPT-4o was prompted to assess each query's dependency on temporal context beyond immediate video boundaries, prioritizing queries whose interpretations or resolutions necessitate referencing historical, user-specific contexts. Examples of selected queries typically involved repeated interactions, habitual activities, or persistent object engagements spanning multiple video sessions.

Following the AI-driven initial screening, we conducted meticulous manual curation to verify query suitability rigorously. Expert annotators reviewed each GPT-4o-identified query, confirming genuine long-context relevance and filtering out queries ambiguous in contextual dependence or insufficiently representative of personalized retrieval demands. Critically, filtered samples required that participants have various candidate video clips, specifically more than 10 videos per participant, with an average candidate set size of approximately 33 clips. This ensured a robust and diverse long-context retrieval environment, capturing rich habitual and episodic nuances.

The resulting refined dataset comprises 245 rigorously filtered videos from 45 participants, providing 639 distinct, carefully validated long-context queries. This comprehensive filtering approach significantly enhances the benchmark's effectiveness in accurately assessing personalized retrieval capabilities under authentic, user-specific episodic memory contexts.

F.1.1 Extended Implementation Details for EgoMemory and EgoRetriever

We leveraged GPT-4o extensively to construct detailed user-specific memory banks by generating comprehensive object-centric metadata annotations from video clips. Additionally, GPT-4o facilitated reflective Chain-of-Thought (CoT) reasoning integral to our retrieval framework. Specifically, we conducted these annotation and reasoning processes for 45 participants, covering a total of 245 videos, over approximately one month. This substantial annotation effort underscores both the computational and temporal investments involved in establishing robust personalized memory contexts.

Our retrieval experiments utilized four NVIDIA V100 GPUs, each equipped with 32GB memory. We evaluated multiple state-of-the-art video-language models, including LanguageBind (Zhu et al., 2024), CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and EgoVLPv2 (Pramanick et al., 2023). Within our proposed EgoRetriever, EgoVLPv2 served as the primary text encoder. To represent videos visually, embeddings from CLIP and BLIP were computed by averaging across embeddings extracted from 15 uniformly sampled frames per video clip. For each candidate video, these visual embeddings were then matched with the GPT-4o-generated textual descriptions using cosine similarity to determine retrieval accuracy. For further detailed methodology and specific parameter settings, please refer to the supplementary materials.

F.2 Broader Impacts

Our proposed framework, EgoRetriever, introduces significant advancements toward personalized, long-context episodic memory retrieval, opening promising avenues for augmented human cognition and improved assistive technologies. However, its deployment also brings potential societal risks warranting careful consideration. First, extensive recording and storage of personalized egocentric data inherently pose substantial privacy concerns, as such continuous visual and contextual capture could inadvertently reveal sensitive personal information or be exploited for unauthorized surveillance and tracking. Secondly, even when functioning correctly, the model might unintentionally reinforce biases embedded in training data or annotations, possibly leading to unequal performance across diverse demographic groups, thus raising fairness considerations. Additionally, misuse of the proposed memory-augmented retrieval technology could facilitate invasive monitoring or targeted manipulation based on personal habits and behaviors, resulting in malicious outcomes such as stalking, identity theft, or psychological manipulation. Incorrect retrieval outcomes, particularly involving sensitive contexts or critical decisions, could also lead to harmful personal or societal consequences. To mitigate these risks, we advocate implementing robust privacy-preserving measures, including data encryption, strict access controls, and user-centric data ownership frameworks. Moreover, comprehensive fairness audits and continuous model eval-

uations across diverse user populations are essential to ensure equitable deployment. Finally, careful consideration of transparent usage policies and developing detection mechanisms for identifying and preventing misuse are crucial steps toward responsibly harnessing the full potential of personalized egocentric video retrieval technologies.

F.3 Extended Related Works

Multimodal Chain-of-Thought Reasoning for Egocentric Video Retrieval. Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable reasoning capabilities, particularly when equipped with Chain-of-Thought (CoT) prompting strategies (Wei et al., 2022; Kojima et al., 2022; Zheng et al., 2023; Zhang et al., 2023). In the domain of composed image retrieval (CIR), Recent advances introduced OSrCIR (Tang et al., 2025b), a training-free, one-stage reflective CoT framework that enables MLLMs to reason about manipulation intent and preserve contextual information, thereby improving retrieval accuracy without the typical information loss associated with two-stage approaches. Similarly, EmbodiedGPT (Mu et al., 2023) extends CoT prompting to vision-language pre-training in embodied environments, leveraging an EgoCOT dataset and prefix-tuned LLMs to enable agents to perform complex action planning through multimodal reasoning. Despite these advances, existing research on episodic memory retrieval from egocentric videos has predominantly focused on short-term or single-video scenarios (Grauman et al., 2022; Hummel et al., 2024; Yang et al., 2025), with limited attention to the long-context, personalized nature intrinsic to human memory. Current benchmarks lack comprehensive personal memory banks and rich, user-specific object annotations (Singh et al., 2016; Damen et al., 2020; Núñez-Marcos et al., 2022; Wang et al., 2023b), which are critical for modeling real-world episodic recall. Building upon recent advances in multimodal CoT reasoning (Tang et al., 2025b; Mu et al., 2023; Mitra et al., 2024b; Zheng et al., 2023; Zhang et al., 2024), our work is the first to bring this paradigm to personalized egocentric video retrieval. We introduce EgoRetriever, a training-free framework that combines MLLMs with reflective CoT prompting, leveraging a comprehensive personal memory bank constructed from extensive user-specific object annotations in Ego4D (Grauman et al., 2022). This enables the model to interpret nuanced user queries and generate detailed, contextually grounded descriptions of target video clips by incorporating recurring personal objects, habitual activities, and social interactions. Extensive experiments on the EgoMemory and EgoCVR (Hummel et al., 2024) benchmarks demonstrate that EgoRetriever consistently outperforms existing baselines, underlining its strong generalizability and promise for real-world deployment in personalized episodic memory retrieval.