

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 MEMORY-AUGMENTED PERSONALIZED RETRIEVAL FOR LONG-CONTEXT EGOCENTRIC VIDEO

Anonymous authors

Paper under double-blind review

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 EgoMemory, EgoCVR, and EgoLifeQA benchmarks 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), which defines personalization as grounding retrieval in the specific objects a user has seen, experienced, or interacted with in humans’ daily life. Recent advances in wearable devices and large language models (LLMs) 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*: (i) data are continuously recorded from the user’s viewpoint; (ii) most queries explicitly reference personal objects (*e.g.*, our empirical analysis in Section 3.2 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 history memory (*e.g.*, “what is the location I play with my dog in last month?”), necessitating models capable of long-context video understanding (*i.e.*, spanning months). This naturally motivates a dynamic personal memory bank that accumulates recurring objects, habits, and social interactions. Since personal interactions repeatedly involve the same items, object frequency is a strong cue of personalized relevance in both intuition and recent works Lee et al. (2012); Lee & Grauman (2015); Grauman et al. (2022); Yang et al. (2025). Enriching user queries with contextualized personal information thus holds significant potential for improving long-context memory retrieval accuracy. Nonetheless, current episodic memory retrieval tasks predominantly concentrate on single-video scenarios Grauman et al. (2022); Hummel et al. (2024) or relatively short-term contexts (*i.e.*, within one week) Yang et al. (2025), neglecting the personalized, long-context nature intrinsic to episodic memory retrieval.

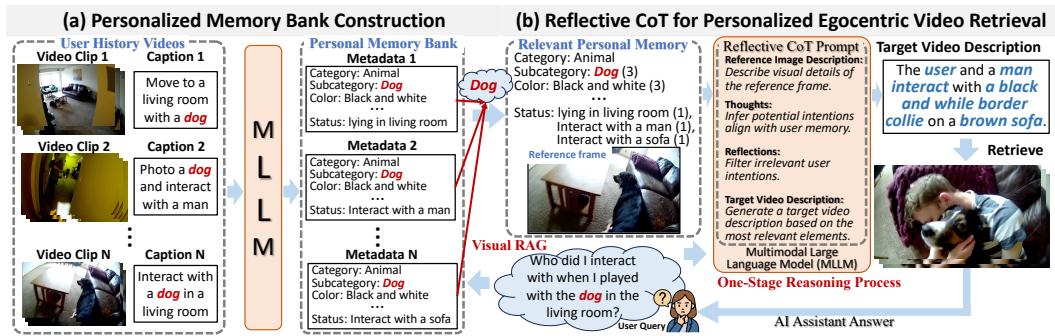


Figure 1: Overview of our approach for personalized egocentric video retrieval, comprising two progressive modules: (a) offline construct a personalized memory bank from each user’s historical videos; (b) online retrieval by query-relevant personal memory to guide intention understanding.

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 propose an annotation pipeline leveraging the advanced reasoning capabilities of MLLMs. Specifically, 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. All candidate queries are further filtered for personalization and long-context via an MLLM-assisted procedure with final human verification ($\sim 91.6\%$ personal, details in Section 4.1). Based on this, we introduce the **EgoMemory** benchmark, designed explicitly for learning personalized information from users’ historical videos to facilitate long-context episodic video retrieval (details in Section 3.2). 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. During inference, the user’s query is analyzed to identify relevant personal objects (e.g., “dog”), after which pertinent memories are retrieved from the personal memory bank to enrich the query and improve subsequent video retrieval precision.

Utilizing the relevant personal memory, we propose **EgoRetriever**, a novel, training-free framework tailored explicitly for long-context episodic video retrieval. As shown in Figure 1(b), **EgoRetriever** combines Multimodal Large Language Models (MLLMs) with a reflective Chain-of-Thought (CoT) prompting strategy to infer nuanced user intentions and generate detailed textual descriptions of target video clips. This approach enhances retrieval accuracy by incorporating fine-grained details, such as the dog’s color and contextual elements like “sofa” and “interaction with a man”, drawn from personal memory.

To summarize, the main contributions are: (1) We introduce a memory-augmented framework for the personalized long-context egocentric video retrieval task and present the **EgoMemory** benchmark, which features individualized memory banks constructed from extensive user-specific object annotations 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 both the EgoMemory and EgoCVR Hummel et al. (2024) and EgoLifeQA Yang et al. (2025) benchmarks demonstrate that **EgoRetriever** consistently and significantly outperforms existing baselines, highlighting its strong generalizability and its potential for real-world deployment in egocentric video retrieval.

2 RELATED WORK

Egocentric Datasets and Benchmarks. Early egocentric studies used ADL Pan et al. (2022), ChadesEgo 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 Ego-ProceL 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

108 multi-view understanding. Recent benchmarks such as EgoSchema Mangalam et al. (2023) and
 109 EgoPlan-Bench Li et al. (2024) emphasize temporal reasoning and planning, while EgoMemoria Ye
 110 et al. (2025) and EgoLife Yang et al. (2025) provide week-long, multi-participant data for study-
 111 ing longer-term behavior. However, these benchmarks generally overlook the fine-grained, person-
 112 specific variability needed for long-context *personalized* retrieval. Our EgoMemory addresses this
 113 gap by offering the first benchmark for personalized egocentric video retrieval, explicitly capturing
 114 inter-individual daily variability and enabling person-centric memory augmentation.

115 **Composed Image and Video Retrieval.** Composed image retrieval (CIR) retrieves images that
 116 are semantically edited by textual prompts Vo et al. (2019); Baldrati et al. (2022). Zero-shot CIR
 117 methods Saito et al. (2023); Baldrati et al. (2023); Tang et al. (2024d); Gu et al. (2024); Karthik
 118 et al. (2024); Tang et al. (2024c); Suo et al. (2024); Du et al. (2024); Tang et al. (2024b; 2025) use
 119 multimodal encoders such as CLIP Radford et al. (2021) to reduce annotation needs, yet often strug-
 120 gle with implicit human intent. Recent training free approaches (e.g., CIReVL Karthik et al. (2024)
 121 and OSrCIR Tang et al. (2024a)) leverage large language models to infer intent and improve compo-
 122 sitional reasoning without supervision. Extending to video, composed video retrieval addresses
 123 temporal complexity. EgoCVR Hummel et al. (2024) supports fine-grained egocentric queries with
 124 a two-stage caption fusion pipeline. Despite progress, current frameworks are still under a model
 125 dynamic context and personal relevance in real egocentric scenarios. We introduce a training-free,
 126 one-stage retrieval framework that grounds user queries in a dynamic personal memory bank and
 127 produces fine-grained, context-aware video descriptions. This design achieves state-of-the-art per-
 128 formance on EgoMemory and advances personal memory retrieval.

129 **Memory Augmented Long Context Retrieval.** Retrieval Augmented Generation (RAG) frame-
 130 works show that coupling large language models with external memory extends reasoning over
 131 long contexts Lewis et al. (2020); Jiang et al. (2023); Shi et al. (2023); Ram et al. (2023); Izac-
 132 ard et al. (2022). Graph augmented retrieval improves multi-hop reasoning by using structured
 133 knowledge graphs built from sources such as Wikipedia and document-level entities for re-ranking
 134 and contextual linking Ding et al. (2019); Zhu et al. (2021); Nie et al. (2019); Das et al. (2019);
 135 Asai et al. (2020); Li et al. (2021). For example, HippoRAG Gutiérrez et al. (2024) employs neu-
 136 robiologically inspired knowledge graphs for advanced reasoning. Lifelogging systems Rossetto
 137 et al. (2020); Nguyen et al. (2021) organize personal data with multimodal knowledge graphs but
 138 are constrained by static schemas and limited flexibility for dynamic, user-driven interpretation.
 139 In contrast, EgoMemory builds a personalized memory bank directly from egocentric video, and
 140 EgoRetriever uses a reflective chain-of-thought prompting within a training-free architecture
 141 to enable dynamic, user-specific reasoning and to produce detailed video descriptions. This yields
 142 significant gains over prior baselines for long context personal retrieval. Additional related work on
 143 multimodal *chain of thought* is provided in the Appendix A.13.

3 METHODOLOGY

144 In this section, we first formalize the proposed long-context video retrieval task with personalized
 145 memory augmentation. We then provide detailed descriptions of the two core components: the
 146 construction of personalized memory banks and the design of the EgoRetriever framework.

3.1 PRELIMINARY

147 We adopt a continuous collection setting where each user records egocentric videos $\mathcal{V} =$
 148 $\{V^{(1)}, \dots, V^{(N)}\}$ over time and contexts. Recordings of varying duration are segmented in postpro-
 149 cessing into semantically coherent clips, yielding the candidate set $\mathcal{C} = \{C_1, \dots, C_M\}$ for retrieval.

150 Personalization requires historical evidence. For each user, we build a memory bank \mathcal{M} with a
 151 pretrained MLLM Ψ_M (Section 3.2). The retrieval task is: given a natural language query Q , retrieve
 152 the most relevant clip $C^* \in \mathcal{C}$. We first parse Q and query \mathcal{M} to obtain personal object metadata
 153 \mathcal{M}_q and a lightweight visual anchor I_r . To avoid conflating scene exposure with personalization,
 154 retrieval into \mathcal{M}_q uses semantic match together with cross-video recurrence and possessive cues.

155 With (Q, \mathcal{M}_q, I_r) , EgoRetriever uses an MLLM with reflective Chain of Thought prompting
 156 to produce a focused target description T_t . A video language retriever (e.g., EgoVLPv2 Pramanick
 157 et al. (2023)) embeds T_t via a text encoder Ψ_T and each candidate C_i via a video encoder Ψ_V . Long

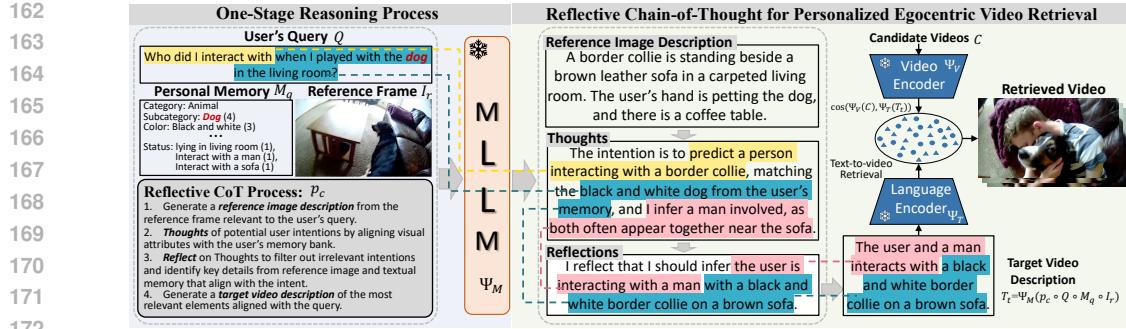


Figure 2: Overview of EgoRetriever. An MLLM processes textual personal memory data M_q , the reference frame I_r , and the user’s query Q to generate the desired target video description T_t by reflective CoT. A vision-language model is then adopted to perform text-to-video retrieval. Texts with different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

context retrieval is performed by ranking candidates with cosine similarity $\cos(\Psi_V(C_i), \Psi_T(T_t))$, returning the top-ranked clips.

3.2 PERSONALIZED MEMORY BANK

Central to personalized retrieval is incorporating long-term, user-specific context rather than incidental scene exposure. From an empirical analysis of Ego4D Grauman et al. (2022) queries using SpaCy Honnibal et al. (2020) and WordNet Miller (1995), **88.4%** explicitly reference physical objects, often linked to the user (e.g., “my bag”, “our dog”). Following prior works Lee et al. (2012); Lee & Grauman (2015); Yang et al. (2025), we operationalize *personalization* as *personally experienced objects*: entities recurring across a user’s videos or linked to the user. To avoid conflating environment priors with personalization, we record cross-video recurrence, within-user attribute consistency, and possessive cues, and down-weight one-off co-occurrences. Specifically, the *Personalized Memory Bank Construction* encodes user-specific object metadata from long-context egocentric video (e.g., attributes, recurrence statistics, and visual exemplars) into a structured memory \mathcal{M} . The *Visual Retrieval-Augmented Generation* then fuses the query with memory entries that are semantically matched (as shown in Figure 1) and supported by cross-video evidence, favoring user-specific contextual cues over incidental context and improving long-context personal video retrieval.

Personalized Memory Bank Construction. We construct the memory bank with a pretrained MLLM Ψ_M ¹, integrating video and narration. In EgoMemory, Ego4D narrations and clips are processed by Ψ_M to extract object attributes (major category, subcategory, texture, shape, color, brand; prompts in Figure 7) and weak personalization cues (first-person/possessive mentions, coreference) together with cross-video recurrence statistics. For example, a *dog* is encoded as *animal→dog* with color/state and recurrence within the same user. Aggregating these profiles across clips yields a structured, user-specific memory \mathcal{M} with entries $\mathbf{m}_i = \Psi_M(C_i)$. Unlike unstructured transcript baselines (e.g., EgoLifeQA Yang et al. (2025)), our representation explicitly models user-specific contextual cues via recurrence and linguistic cues, helping reduce the weight of incidental scene exposures. Practically, since egocentric capture includes idle periods (e.g., sleep), construction can run asynchronously with minimal impact on interactive use.

Visual Retrieval-Augmented Generation. The memory bank serves as a semantic repository of personal experiences. Given a query Q , **Visual RAG** proceeds: (1) extract the object subcategory and retrieve $\mathcal{M}_q \subset \mathcal{M}$ using semantic match and recurrence/possessive cues (to reduce environment priors); (2) summarize attribute distributions in \mathcal{M}_q into compact textual context; (3) pick a reference frame I_r by selecting a reference video V_r (centroid of the retrieved set) and taking its middle frame. I_r is a lightweight visual anchor providing historical context and need not appear in the target clip C^* . Reflective CoT prompting in EgoRetriever then reasons over plausible evolutions (object/state/location), producing a focused description T_t that prioritizes user-linked cues over incidental context. Thus, the enriched tuple (Q, \mathcal{M}_q, I_r) improves personalized long-context video retrieval. For details, please refer to Appendix A.1.

¹Efficiency aspects (e.g., incremental extraction/updating) are orthogonal and deferred to future work.

216 3.3 REFLECTIVE CHAIN-OF-THOUGHT FOR PERSONAL EGOCENTRIC VIDEO RETRIEVAL
217

218 Conventional egocentric retrieval (*e.g.*, TFR CVR) uses a two-stage pipeline to form the target de-
219 scription, which can lose visual detail and dilute user-specific cues. We introduce EgoRetriever,
220 a one-stage, training-free framework for long context personal retrieval. It leverages an MLLM to
221 directly produce a detailed target description conditioned on the query, personal memory, and a
222 lightweight visual anchor, without additional training. Formally, with MLLM Ψ_M ,

$$223 \quad 224 \quad T_t = \Psi_M(p_c \circ Q \circ \mathcal{M}_q \circ I_r). \quad (1)$$

225 The reflective CoT prompt p_c (Figure 2) guides single prompt reasoning with three concise stages
226 (full template in Appendix A.2.2):
227

228 **Reference Image Description.** During this initial step, the MLLM provides a detailed description
229 of the visual content relevant to the user’s query. In Figure 2, irrelevant elements such as general
230 room features (*e.g.*, coffee table) are selectively omitted, while intention-relevant elements (*e.g.*, a
231 black and white border collie standing beside a brown leather sofa, user’s hand petting the dog) are
232 preserved to clearly align with the user’s retrieval intent.
233

234 **Thoughts.** Given the relevant visual details and the user’s query, the MLLM interprets the implicit
235 retrieval intent. The MLLM explicitly reasons about which visual and contextual cues most sig-
236 nificantly influence its understanding of the user’s query. Specifically, it identifies critical visual
237 attributes (*e.g.*, the black and white border collie) and contextual information from the personalized
238 memory bank (*e.g.*, frequent interaction involving a man and a sofa). This reasoning aligns the visual
239 cues with memory patterns, guiding the inference toward the user’s probable interaction partner.
240

241 **Reflections.** Given the potential intentions and reference image, the MLLM filters the inferred
242 intentions by explicitly considering the coherence and context of visual and textual details. Potential
243 irrelevant assumptions (*e.g.*, interactions unrelated to the query context) are excluded. The MLLM
244 clarifies the rationale behind identifying a man interacting with the user and the border collie near a
245 brown sofa, thereby reducing ambiguity and hallucinations.
246

247 **Target Video Description.** Finally, given the filtered intentions and relevant visual-contextual rea-
248 soning results, the MLLM generates the accurate target video description. This description suc-
249 cinctly captures the user’s intended interaction (*e.g.*, the user and a man interacting with the black
250 and white border collie on a brown leather sofa), clearly matching the user’s original retrieval query.
251

252 After generating the target description T_t , EgoRetriever uses a video-language retrieval back-
253 bone (*e.g.*, EgoVLPv2 Pramanick et al. (2023)) to identify the most relevant video clips from a
254 candidate pool. T_t is encoded using a pretrained text encoder Ψ_T , and each candidate clip C_i is em-
255 bedded using a video encoder Ψ_V . The most relevant clip C^* is then obtained via cosine similarity:
256

$$257 \quad 258 \quad C^* = \operatorname{argmax}_{C_i \in \mathcal{C}} \frac{\Psi_V(C_i)^\top \Psi_T(T_t)}{\|\Psi_V(C_i)\| \|\Psi_T(T_t)\|}. \quad (2)$$

259 4 EXPERIMENTS
260261 4.1 EGOMEMORY BENCHMARK
262

263 To rigorously evaluate *personalized* egocentric video retrieval, we introduce the **EgoMemory**
264 benchmark, derived from Ego4D’s Natural Language Queries (NLQ) Grauman et al. (2022). NLQ
265 provides ~ 227 hours of head-mounted video from 137 participants across 74 locations, annotated
266 with free-form queries about “when/where/with whom/what,” reflecting realistic recall scenarios.
267

268 **Limitations of Ego4D’s NLQ for Personalization.** While rich and diverse, NLQ was designed
269 for temporal localization within *individual* clips and does not aggregate multi-video, user-specific
270 context. In particular, videos are not grouped by user at retrieval time, ownership or user-linkage
271 cues are not modeled, and queries need not require long-horizon evidence. Consequently, evaluating
272 memory-augmented personalized retrieval directly on NLQ risks rewarding scene/frequency priors
273 rather than genuine personalization.
274

270 **Design of EgoMemory: user grouping and query filtering.** To address this, we treat each user as a distinct retrieval
 271 unit by aggregating *all* of their videos as personal context at
 272 query time. From 137 participants, we manually select 45
 273 with sufficient temporal coverage to emulate practical AR us-
 274 age. To ensure that evaluation genuinely probes personaliza-
 275 tion *and* long-context reasoning, we apply a two-stage filtering
 276 pipeline: (i) **GPT-4o pre-screening** to keep queries with ex-
 277 plicit personal references or strong user linkage (*e.g.*, posses-
 278 sives, deictic cues) and to discard queries designed for short
 279 single-clip answers; (ii) **Human verification** to confirm that
 280 the main referenced object is plausibly user-linked across his-
 281 tory. Concretely, for each retained query–target pair, annota-
 282 tors review 20 additional short clips from the same user con-
 283 taining the same object class; queries are labeled *personal*
 284 when $\geq 90\%$ of reviewed instances match the target and *un-*
 285 *certain* when $\geq 75\%$. This results in $\sim 91.6\%$ personal queries.
 286 The resulting dataset comprises 245 videos from 45 users and
 287 639 queries. For details, please refer to the Appendix A.3.

Memory banks and added annotations. For each participant, we annotate 165,795 user-specific object annotations across 12 attributes (category, color, texture, shape, brand, state, etc.) for personal memory bank construction (Section 3.2). The number of unique object types per user ranges from 59 to 638 (median 129); memory sizes range from 322 to 10,454 entries (median 1,312). This structured representation goes beyond unstructured transcripts by explicitly modeling user-specific contextual cues (*i.e.*, recurrence + linguistic cues) and supports training-free, interpretable personalization.

Personalization Heterogeneity. To quantify user specificity, we compute Jaccard similarity over attribute sets for the 100 most frequent object types, comparing inter- vs. intra-participant distributions. Also, to account for environment bias, we recompute inter-participant similarity, restricting comparisons to matched coarse scenes (kitchen, living room, outdoor) using the “Status” metadata, i.e., $J(A^{\text{scene}=s}, B^{\text{scene}=s}) = \frac{|A \cap B|}{|A \cup B|}$. As shown in Figure 4, **68.3%** of objects < 0.4 , confirming that personalization persists beyond scene priors.

Candidate set for retrieval. The candidate retrieval pool contains 2,228 clips from participants' histories (mean ~ 33 clips/user), spanning 4 to 300 s (mean 103.82 s), totaling 64.25 h. By construction, selected queries target objects recurring across a user's videos, so relevant evidence often lies outside the target clip. This constitutes a *long-context* setting that requires integrating (Q, \mathcal{M}_q, I_r) with cross-video evidence rather than relying on single-clip shortcuts.

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 A.9.

Implementation Details. We leverage GPT-4o for constructing the user-specific memory banks by generating detailed object-centric metadata from video clips. Additionally, GPT-4o is also employed to perform reflective CoT reasoning. Our retrieval experiments were conducted using four NVIDIA V100 GPUs with 32GB each. We evaluated multiple state-of-the-art video-language models, including LanguageBind Zhu et al. (2023), CLIP Radford et al. (2021), BLIP Li et al. (2022), and EgoVLPv2 Pramanick et al. (2023). We employ EgoVLPv2 as the text encoder in EgoRetriever. CLIP and BLIP visual representations for the videos are obtained by averaging embeddings from 15 uniformly sampled image frames. For each candidate video, visual embeddings were extracted and subsequently matched with the GPT-4o-generated textual descriptions via cosine similarity. Please refer to the Appendix A.10 for more details.

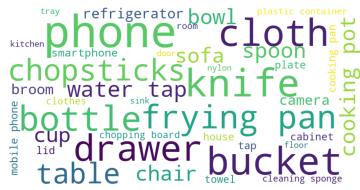


Figure 3: The most frequently objects in constructed memory banks.

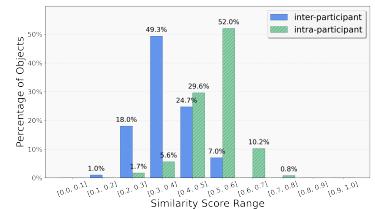


Figure 4: Similarity of top 100 objects in constructed memory banks.

324
 325 Table 1: Mean Recall@K (%) for different retrieval configurations. “Video Model” indicates use of
 326 a video encoder, “Textual Memory Bank” denotes personal text-based memory, “Visual Reference”
 327 represents visual info in the reference image, and “Fusion Strategy” specifies modality combination
 (“Avg” refers to naive average fusion). The best and second-best results are in **bold** and underlined.
 328

329 330 Method	331 Video Model	332 Textual Memory Bank	333 Visual Reference	334 Fusion Strategy	335 Mean Recall (%) 336 mR@1 mR@2 mR@3		
	337 X	338 ✓	339 X	340 —	341 3.62	342 9.74	343 15.23
Random	X	✓	X	—	10.41	12.95	16.72
CLIP	X	✓	X	—	10.88	13.67	17.48
BLIP	X	✓	X	—	11.25	15.03	18.30
EgoVLPv2	✓	✓	X	—	11.02	14.60	17.83
LanguageBind	✓	✓	X	—			
CLIP	X	X	✓	Avg	14.74	17.53	21.41
BLIP	X	X	✓	Avg	15.12	18.07	22.82
EgoVLPv2	✓	X	✓	Avg	15.77	20.61	23.79
LanguageBind	✓	X	✓	Avg	15.26	20.14	23.04
CLIP	X	✓	✓	Avg	15.64	18.63	22.71
BLIP	X	✓	✓	Avg	16.02	19.17	24.12
EgoVLPv2	✓	✓	✓	Avg	16.67	21.71	25.09
LanguageBind	✓	✓	✓	Avg	16.16	21.24	24.34
BLIP _{CoVR} Ventura et al. (2024)	X	✓	✓	Cross-Attn.	15.94	19.17	23.00
BLIP _{CoVR-ECDE} Thawakar et al. (2024)	X	✓	✓	Cross-Attn.	16.41	19.63	23.64
CIReVL Karthik et al. (2024)	X	✓	✓	Captioning	16.95	20.13	24.37
OSrCIR Tang et al. (2024a)	X	✓	✓	Captioning	17.28	21.64	25.49
TFR-CVR Hummel et al. (2024)	✓	✓	✓	Captioning	<u>18.21</u>	<u>27.12</u>	<u>32.05</u>
EgoRetriever (Ours)	✓	✓	✓	Captioning	23.19	38.48	47.83

344 4.2 MAIN RESULTS

345
 346 We compare EgoRetriever to three families of egocentric retrieval systems: (i) *Training free*
 347 *encoders*, **CLIP** Radford et al. (2021), **BLIP** Li et al. (2022), **EgoVLPv2** Pramanick et al. (2023),
 348 and **LanguageBind** Zhu et al. (2023), with video features from 15 uniformly sampled frames, eval-
 349 uated under three input regimes: (a) query plus textual memory, (b) query plus visual reference,
 350 and (c) late fusion of all three. (ii) *Composed image retrieval*, **BLIP_{CoVR}** Ventura et al. (2024),
 351 **BLIP_{CoVR-ECDE}** Thawakar et al. (2024), **CIReVL** Karthik et al. (2024), and **OSrCIR** Tang et al.
 352 (2024a), which generate a target description with an LLM and retrieve, typically via CLIP. (iii) *Ego-*
 353 *centric aware TFR-CVR* Hummel et al. (2024), which captions a key frame and prompts an LLM
 354 to form the target description before first-person video retrieval. Unlike these two-stage pipelines,
 355 EgoRetriever performs one-stage reflective reasoning within an MLLM to directly produce the
 356 target description. Both TFR-CVR and EgoRetriever use the same MLLM (GPT 4o), and all
 357 baselines receive identical inputs and encoders, ensuring fair comparison.

358 Table 1 reports mean Recall@K performance for various retrieval configurations on the EgoMem-
 359 ory benchmark. Our EgoRetriever achieves the best performance across all metrics, with a
 360 notable mR@1 of 23.19% and mR@3 of 47.83%, outperforming the best egocentric-aware base-
 361 line (TFR-CVR) by 4.98% and 15.78%, respectively. This significant improvement underscores
 362 the effectiveness of our reflective CoT prompting and personalized memory bank design for long-
 363 context, user-centric retrieval. Ablation across modality configurations demonstrates that models
 364 relying solely on textual memory are limited in capturing user intent (e.g., EgoVLPv2 mR@1:
 365 11.25%), highlighting the necessity of integrating visual references. The combination of textual
 366 memory and visual reference leads to consistent gains, with our approach delivering the highest
 367 recall even in complex, diverse scenarios. Compared to state-of-the-art composed image retrieval
 368 models, EgoRetriever yields a substantial relative gain, improving mR@1 by 5.91% over OS-
 369 rCIR and nearly doubling mR@3 performance. These results validate the value of leveraging rich,
 370 user-specific historical data and long-context modeling: by constructing comprehensive personal
 371 memory banks from aggregated user histories (*i.e.*, with thousands of annotated attributes per user),
 372 our framework enables long-context reasoning over user context, habits, and object interactions that
 373 are critical for accurate retrieval. For more qualitative analysis, please refer to the Appendix A.2.4.

374 4.3 ABLATION STUDY AND PERFORMANCE ANALYSIS

375
 376 In Table 2, we assess the contribution of each component on EgoMemory. **(1) Models ‘2–7’ eval-**
 377 **uate 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

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 2: Results in terms of mR@1, mR@2, and mR@3 demonstrating the necessity of the key modules in EgoRetriever.

Methods	mR@1	mR@2	mR@3
1. Full Model (GPT-4o)	23.19	38.48	47.83
Significance of key modules of our EgoRetriever			
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
Impact of different CoT methods			
8. Simple CoT	20.13	33.18	41.28
9. Advance CoT	19.42	32.50	41.02
Practical feasibility without human-written narrations			
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: Results that emphasize the importance of our personal memory bank.

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

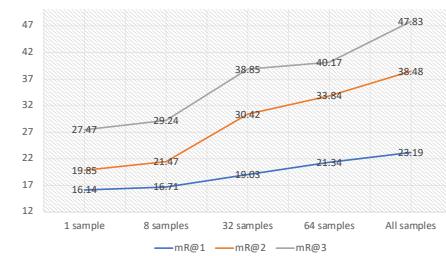


Figure 5: Effect of the number of metadata for each object in the memory bank.

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) Models ‘8–9’ compare our Reflective CoT against other CoT methods.** Replacing our reflective CoT with simple CoT or an advanced two-stage CoT (DDCoT) degrades mean Recall by about 5%, 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 (model ‘12’) show only minor declines, confirming that EgoRetriever remains effective without ground truth narrations and is practical for real-world deployment. **(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 promising balance between efficiency and retrieval performance.

4.4 ANALYSIS

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

Analysis of Memory Bank Design. Table 3 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.

Generalization ability of our EgoRetriever. We evaluate EgoRetriever on EgoCVR Hummel et al. (2024), which similarly requires generating a target video description from a reference image and a textual edit; following protocol, we report both *global* and *local* search. As shown in Table 4, EgoRetriever achieves the best results in all settings, surpassing the two stage TFR-CVR*

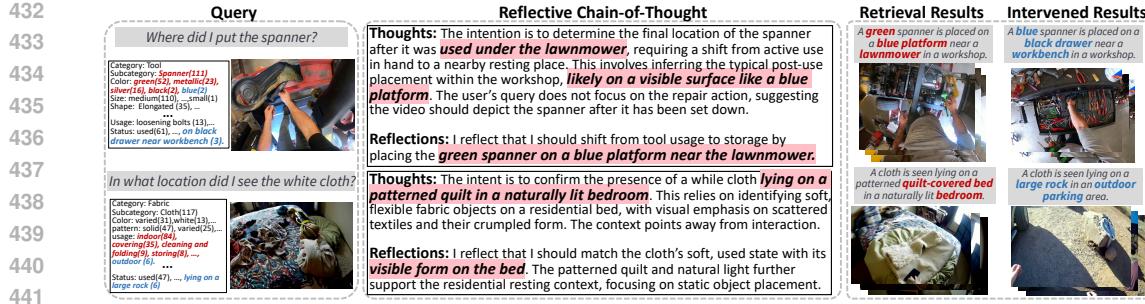


Figure 6: Visualization of common failure cases. The red color denotes the incorrect reasoning outcomes of intention. The top-1 retrieval result and the intervened correction are shown.

baseline (GPT-4o as captioner and LLM) by an average of 5.05%. We further test transfer to EgoLifeQA Yang et al. (2025), a fixed answer QA suite probing entity logs, past event recall, and habit patterns, where EgoRetriever obtains SoTA, with average performance gains of 6.17% over the EgoGPT. Together, these results demonstrate strong generalization of our one-stage, training-free framework beyond EgoMemory. For details, please refer to the Appendix A.4,A.5.

Impact of Memory Bank Context Length. We evaluate the influence of memory bank context length by varying the number of included object metadata, as shown in Figure 5. Our results reveal that restricting the memory bank to short-term contexts significantly limits retrieval performance, as essential long-term user patterns and object interactions are underrepresented. As the context length increases, incorporating a broader history of user experiences, retrieval accuracy improves markedly. Notably, substantial performance gains are observed when extending the memory bank to encompass more metadata, highlighting the importance of long-context information for modeling fuser semantics. These findings underscore that a comprehensive, extensive memory bank is crucial for enabling accurate and personalized long-context video retrieval.

Analysis of Common Failure Cases. To assess the limitations of EgoRetriever, we examined 100 failure cases. As shown in Figure 6, we identify two primary issues: (1) *Object Disambiguation Challenges* (74%): The model often fails to accurately distinguish target objects in cluttered scenes, such as identifying a specific spanner among visually similar tools in a workshop (Row 1). (2) *Context Misinterpretation* (21%): The model may misinterpret user intent when the visual reference context is ambiguous. For example, retrieving an indoor scene for a “white cloth” when the correct context is outdoors (Row 2). Notably, supplementing queries with more detailed contextual cues (e.g., *black drawer*, *outdoor*) can mitigate these errors, highlighting the need for enhanced object differentiation and context reasoning in personalized, long-context egocentric video retrieval.

5 CONCLUSION

In this paper, we tackle the challenge of personalized, long-context episodic memory retrieval from egocentric video. We introduce EgoMemory, a benchmark built from Ego4D with user-specific memory banks and diverse, context-rich queries. We further propose EgoRetriever, a training-free framework leveraging Multimodal Large Language Models and reflective Chain-of-Thought reasoning to explicitly understand user queries through personal memory for personalized video retrieval. Extensive experiments on EgoMemory and EgoCVR demonstrate that our approach achieves state-of-the-art performance and strong generalization, marking a significant advance in practical personalized and long-context egocentric video retrieval. It inspires future research on user-centric memory augmentation and has broad implications for real-world multimodal AI applications.

486
487
ETHICS STATEMENT488
489
490
491
Scope and alignment with the ICLR Code of Ethics. Our work follows the ICLR principles
of responsible stewardship: contributing to well-being, upholding scientific excellence, avoiding
harm, being honest and transparent, ensuring fairness and non-discrimination, respecting prior work,
respecting privacy, and honouring confidentiality.492
493
494
495
Human data, consent, and provenance. All experiments use public egocentric datasets (Ego4D,
EgoCVR, EgoLifeQA) that were collected under their own consent and governance processes. We
introduce no new human data collection and do not attempt re-identification or linkage to external
records. We comply with dataset licenses and intended use policies and acknowledge all sources.496
497
498
499
500
Privacy by design. Our method is training-free and centers on personalization of the device. The
personal memory bank is generated and stored locally during idle periods, giving users control over
creation, inspection, and deletion. At query time, only minimal structured metadata (attribute tuples
and frequency counts) is shared with the language model. Raw video, audio, faces, names, locations,
and other directly identifying content are not transmitted. Open source MLLMs can replace hosted
services for fully local inference when stricter privacy is required.501
502
503
504
505
Transparency, reproducibility, and integrity. We report methods, prompts, model choices, and
evaluation protocols to support replication. We do not fabricate or obfuscate results, and we dis-
close limitations and failure modes. If released, code and de-identified annotations will include
documentation of data provenance and usage constraints.506
507
508
509
510
Fairness and potential harms. Egocentric data can encode social and environmental biases. Our
memory construction focuses on object-level attributes rather than demographic attributes, and we
report per benchmark generalization. We discourage use in surveillance or monitoring of indi-
viduals. Any released resources will carry a license that prohibits re-identification, profiling, law
enforcement, or use targeting protected classes.511
512
513
514
Confidentiality and data handling. We do not handle confidential third-party data. When hosted
MLLMs are used, we enable no retention settings where available and minimize payloads. Access
credentials are stored outside the artifacts released for research.515
516
517
Annotator well-being. Manual verification was performed by trained team members following
internal guidelines. Annotators could skip any item, and no harmful content was intentionally intro-
duced.518
519
520
521
Environmental impact. We emphasize efficiency through precomputed embeddings and single-
stage reasoning. Measured query latency and memory remain stable as data volume grows, which
reduces energy cost and enables deployment on modest hardware.522
523
524
Limitations and remediation. A formal privacy guarantee, such as differential privacy, is out of
scope and is an important direction for future work. We welcome community feedback and will
follow ICLR processes for raising concerns and remediation if any ethical issues are identified.525
526
REPRODUCIBILITY STATEMENT527
528
529
530
531
532
533
534
535
536
537
538
539
We took several steps to ensure that our results are reproducible. An anonymized repository included
in the supplementary materials provides training/inference scripts, configuration files, environment
specifications, and pretrained checkpoints for the backbones used in our experiments, along with
a minimal working example and sample data to verify end-to-end execution. The full method is
specified in the main text (Section 3.4), with the complete algorithmic procedure and prompt tem-
plates given in Appendix A.2.1–A.2.3 and qualitative analyses in A.2.4; the personalized memory
bank design, construction template, and cross-user similarity algorithm are detailed in A.1.1–A.1.2
with additional qualitative analysis in A.1.3. Dataset usage and evaluation details are organized per
benchmark: EgoMemory (A.3), EgoCVR (A.4), and EgoLifeQA (A.5), including task definitions,
preprocessing, official protocols, and metric definitions. We report ablations and sensitivity studies
in A.6, and provide compute/runtime profiling and latency breakdowns in A.7–A.8. Candidate set
statistics are summarized in A.9, and further implementation notes (e.g., seed control, batch-
ing, precision settings) appear in A.10. To sum, these materials are intended to allow independent
researchers to reproduce the key results reported in the paper with minimal additional assumptions.

540 REFERENCES
541

542 Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning
543 to retrieve reasoning paths over wikipedia graph for question answering. In *International
544 Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJgVHkrYDH>.

545

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

549

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

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

554

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

556

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

561

562 Rajarshi Das, Ameya Godbole, Dilip Kavarthapu, Zhiyu Gong, Abhishek Singhal, Mo Yu, Xi-
563 aoxiao Guo, Tian Gao, Hamed Zamani, Manzil Zaheer, and Andrew McCallum. Multi-step
564 entity-centric information retrieval for multi-hop question answering. In Adam Fisch, Alon
565 Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen (eds.), *Proceedings of the
566 2nd Workshop on Machine Reading for Question Answering*, pp. 113–118, Hong Kong, China,
567 November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5816. URL
<https://aclanthology.org/D19-5816>.

568

569 Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive graph for multi-hop
570 reading comprehension at scale. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.),
571 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.
572 2694–2703, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/
573 v1/P19-1259. URL <https://aclanthology.org/P19-1259>.

574

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

577

578 Yue Fan, Xiaojian Ma, Ruijie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A
579 memory-augmented multimodal agent for video understanding. In *European Conference on Com-*
580 *puter Vision*, pp. 75–92. Springer, 2024.

581

582 Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Gird-
583 har, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in
584 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision
585 and pattern recognition*, pp. 18995–19012, 2022.

586

587 Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos
588 Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, et al. Ego-exo4d: Understanding
589 skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF
590 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

591

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

595

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

594 Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-
 595 strength natural language processing in python. 2020.

596

597 Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong
 598 Lu, Yali Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view
 599 of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer
 600 Vision and Pattern Recognition*, 2024.

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

604

605 Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A.
 606 Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval aug-
 607 mented language models. *ArXiv*, abs/2208.03299, 2022. URL <https://arxiv.org/abs/2208.03299>.

608

609 Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang,
 610 Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor,
 611 Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Meth-
 612 ods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Associa-
 613 tion for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL <https://aclanthology.org/2023.emnlp-main.495>.

614

615 Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-
 616 language for training-free compositional image retrieval. In *The Twelfth International Confer-
 617 ence on Learning Representations*, 2024. URL <https://openreview.net/forum?id=EDPxjXzSb>.

618

619 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
 620 language models are zero-shot reasoners. *Advances in neural information processing systems*,
 621 35:22199–22213, 2022.

622

623 Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summariza-
 624 tion. *International Journal of Computer Vision*, 114(1):38–55, 2015.

625

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

629

630 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 631 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe
 632 Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the
 633 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook,
 634 NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL <https://dl.acm.org/doi/abs/10.5555/3495724.3496517>.

635

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

639

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

643

644 Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and
 645 Bingquan Liu. Hopretriever: Retrieve hops over wikipedia to answer complex questions. *Pro-
 646 ceedings of the AAAI Conference on Artificial Intelligence*, 35:13279–13287, 05 2021. doi:
 647 10.1609/aaai.v35i15.17568.

648

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

648 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
 649

650 Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagno-
 651 stic benchmark for very long-form video language understanding. In *Thirty-seventh Confer-
 652 ence on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL
 653 <https://openreview.net/forum?id=JV1Wseddak>.

654 George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):
 655 39–41, 1995.
 656

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

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

663 Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhui Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng
 664 Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of
 665 thought. *Advances in Neural Information Processing Systems*, 36:25081–25094, 2023.
 666

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

670 Yixin Nie, Songhe Wang, and Mohit Bansal. Revealing the importance of semantic retrieval for
 671 machine reading at scale. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.),
 672 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and
 673 the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.
 674 2553–2566, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:
 675 10.18653/v1/D19-1258. URL <https://aclanthology.org/D19-1258>.

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

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

682 Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou,
 683 Rama Chellappa, and Pengchuan Zhang. Egolpv2: Egocentric video-language pre-training with
 684 fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer
 685 Vision*, pp. 5285–5297, 2023.

687 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
 688 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
 689 Sutskever. Learning transferable visual models from natural language supervision. In *Proceed-
 690 ings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

691 Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and
 692 Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association
 693 for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL <https://aclanthology.org/2023.tacl-1.75>.

695 Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernischov, and
 696 Abraham Bernstein. Lifograph: a knowledge graph for lifelogs. In *Proceedings of the Third
 697 Annual Workshop on Lifelog Search Challenge*, pp. 13–17, 2020.

698 Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas
 699 Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Pro-
 700 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19305–
 701 19314, 2023.

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

706 Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke
 707 Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language
 708 models. *ArXiv*, abs/2301.12652, 2023. URL <https://api.semanticscholar.org/CorpusID:256389797>.

710 Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari.
 711 Charades-ego: A large-scale dataset of paired third and first person videos, 2018. URL <https://arxiv.org/abs/1804.09626>.

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

717 Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequenti-
 718 al recommendation with bidirectional encoder representations from transformer. In *Proceedings
 719 of the 28th ACM international conference on information and knowledge management*, pp. 1441–
 720 1450, 2019.

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

725 Yuanmin Tang, Xiaoting Qin, Jue Zhang, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Ling,
 726 Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflec-
 727 tive chain-of-thoughts for training-free zero-shot composed image retrieval. *arXiv preprint
 728 arXiv:2412.11077*, 2024a.

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

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

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

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

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

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

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

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

756 Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering
 757 queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023b.
 758

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

762 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 763 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
 764 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
 765 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,
 766 Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
 767 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint*
 768 *arXiv:2412.15115*, 2024.

769 Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang,
 770 Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life
 771 assistant. *arXiv preprint arXiv:2503.03803*, 2025.

772 Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang,
 773 Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MMEgo: Towards building ego-
 774 centric multimodal LLMs for video QA. In *The Thirteenth International Conference on Learning*
 775 *Representations*, 2025. URL <https://openreview.net/forum?id=67sSPPAZiG>.
 776

777 Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo.
 778 Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image
 779 inputs, 2024. URL <https://arxiv.org/abs/2401.02582>.

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

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

786 Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang,
 787 Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-
 788 modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

789 Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. Adaptive information
 790 seeking for open-domain question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia
 791 Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods*
 792 *in Natural Language Processing*, pp. 3615–3626, Online and Punta Cana, Dominican Republic,
 793 November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.
 794 293. URL <https://aclanthology.org/2021.emnlp-main.293>.
 795

796 A APPENDIX

799 TABLE OF CONTENTS FOR APPENDIX

801 • A.1 Details for Our Personalized Memory Bank	16
802 – A.1.1 Complete Template for Personalized Memory Bank Construction	16
803 – A.1.2 Algorithm of Calculating Cross-User Object Attribute Similarity	18
804 – A.1.3 Qualitative Analysis for Personalized Memory Bank	18
805 • A.2 Details for Our EgoRetriever’s Process	19
806 – A.2.1 Algorithm of EgoRetriever’s Process	19
807 – A.2.2 Complete Template for Reflective CoT	20
808 – A.2.3 Vision-by-Language In-Context Learning Details	21

810	– A.2.4 Qualitative Analysis of Reflective CoT	25
811	• A.3 Details of EgoMemory Benchmark	25
812	• A.4 Experiment Details on EgoCVR Benchmark	27
813	– A.4.1 EgoCVR Benchmark Overview	27
814	– A.4.2 Task Definition	27
815	– A.4.3 Dataset Construction	27
816	– A.4.4 Evaluation Protocol	27
817	– A.4.5 Evaluation Metrics	28
818	– A.4.6 Experimental Setup	28
819	– A.4.7 Results	28
820	– A.4.8 Comparison of Input Modalities: Our Benchmark vs. EgoCVR	28
821	• A.5 Experiment Details on EgoLifeQA Benchmark	29
822	– A.5.1 EgoLifeQA Benchmark Overview	29
823	– A.5.2 Task Definition	29
824	– A.5.3 Dataset Construction	29
825	– A.5.4 Evaluation Protocol	30
826	– A.5.5 Evaluation Metrics	30
827	– A.5.6 Experimental Setup	30
828	– A.5.7 Results	30
829	• A.6 More Ablation Study on EgoRetriever and reference frame selection	30
830	• A.7 Detailed Comparison of Computational Cost	31
831	• A.8 Scalability under Increasing Video/Metadata Volume	31
832	• A.9 Candidate Set Statistics	32
833	• A.10 More Implementation Details	33
834	– A.10.1 Details of Rigorous Data Filtering	33
835	– A.10.2 Extended Implementation Details for EgoMemory and EgoRetriever ..	33
836	• A.11 Limitation	33
837	• A.12 Broader impacts	34
838	• A.13 More Related Works	34
839	• A.14 The Use of Large Language Models (LLMs)	35
840		
841		
842		
843		
844		
845		
846		
847		
848		
849		
850		
851		
852		
853		
854		
855		
856		
857		
858		
859		
860		
861		
862		
863		

A.1 DETAILS FOR OUR PERSONALIZED MEMORY BANK

A.1.1 COMPLETE TEMPLATE FOR PERSONALIZED MEMORY BANK CONSTRUCTION

In constructing the personalized memory bank, we leverage a structured prompt that guides a pretrained multimodal large language model (MLLM) to systematically extract detailed object attributes from textual narrations and visual contexts, as outlined in Figure 7. 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.

864 You are a helpful vision assistant that identifies object attributes in a structured JSON format.
 865 Ensure your output is valid JSON with exactly the specified top-level keys.
 866
 867 Please carefully analyze the following sentence and identify the primary object being interacted
 868 with by the speaker.
 869 The sentence is: "{object_sentence}"
 870
 871 Consider both the sentence and, if applicable, any associated visual information to determine the
 872 object and its attributes.
 873
 874 Your task is to describe this object's attributes in a structured JSON format.
 875 The JSON output must contain EXACTLY the following top-level keys:
 876 {required_keys_str} # Dynamically populated list, e.g., major_category, subcategory, color, texture,
 877 shape, material, brand, style, pattern, feature, usage, status.
 878
 879 ## Guidance for attribute values
 880 1. Be Specific and Descriptive: For each attribute, provide the most accurate and detailed value you
 881 can infer. For example, for 'color', if an object is "dark blue with yellow stripes", please state that
 882 rather than just "blue" or "patterned".
 883 2. Use Known Values as Examples: Below is a list of attribute categories and examples of values
 884 seen previously. Use these to understand the type of information expected. If a relevant value is
 885 present, you can use it.
 886 Known attribute examples:
 887 {known_str_joined} # Dynamically populated, e.g., "- color: (e.g., red, blue, green, ...)"
 888 3. Invent New Values When Necessary: If the object has a characteristic not covered by the
 889 examples or if the examples are not relevant, provide a new, concise, and descriptive value. This is
 890 how we discover new attributes.
 891 4. Handling Uncertainty/Not Applicable: If an attribute's value cannot be determined from the
 892 provided information or is not applicable to the object, use "unspecified" or "N/A". Avoid guessing
 893 if confidence is very low. For 'brand', if not explicitly mentioned or visible, "unspecified" is
 894 appropriate.
 895 5. Compound Attributes: For attributes like 'material' (e.g., "fabric and wood") or 'feature' (e.g.,
 896 "supporting and sleeping"), list the distinct components as a single string.
 897 6. Consistency: Ensure your entire output is a single, valid JSON object with all string values
 898 properly escaped.
 899
 900 Example of desired thinking for 'color' if a bed is blue and white patterned:
 901 "color": "blue and white patterned"
 902
 903 Now, based on the sentence "{object_sentence}" and any visual context, provide the JSON output.

Figure 7: The complete template of our personal memory construction.

904
 905
 906
 907
 908 **Specificity and Descriptive Precision.** To maximize the accuracy and richness of the memory
 909 bank, the prompt instructs the model to provide detailed, descriptive attribute values. It explicitly
 910 discourages vague descriptions, advocating specificity, for instance, specifying "dark blue with yel-
 911 low stripes" rather than a generic label like "blue" or "patterned." This precision enhances the utility
 912 of the memory bank for fine-grained retrieval.
 913

914 **Known Values and Inventive Flexibility.** The prompt includes dynamically populated examples of
 915 known attribute values, providing clarity and consistency in expected responses. However, recogniz-
 916 ing the inherent novelty in egocentric video contexts, the prompt encourages the model to introduce
 917 new, concise, and descriptive attribute values when existing examples are insufficient, thereby con-
 918 tinually enriching the attribute taxonomy.

918 **Algorithm 1** Calculating Average Diversity for an Object Class
919 **Input:** An object class O , a global attribute database \mathcal{D} mapping each user U_k to their attribute set
920 A_k for object O .
921 **Output:** Average diversity score $\text{Div}_{\text{avg}}(O)$ for object class O .
922 1: 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 .
923
924 2: Initialize a list of similarity scores $S = []$.
925 3: **for** each unique pair of users (U_i, U_j) in \mathcal{U}_O where $i \neq j$ **do**
926 4: Retrieve attribute sets for each user: $A_i = \mathcal{D}(U_i, O)$ and $A_j = \mathcal{D}(U_j, O)$.
927 5: Compute the Jaccard similarity:
928
929
$$s_{ij} = J(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

930
931 6: Add s_{ij} to S .
932 7: **end for**
933 8: **if** S is not empty **then**
934 9: Compute the average similarity: $\text{Sim}_{\text{avg}}(O) = \frac{1}{|S|} \sum_{s \in S} s$
935 10: Compute the average diversity: $\text{Div}_{\text{avg}}(O) = 1 - \text{Sim}_{\text{avg}}(O)$
936 11: **else**
937 12: Set average diversity to 0 (or undefined if only one user has the object): $\text{Div}_{\text{avg}}(O) = 0$
938 13: **end if**
939 14: **return** $\text{Div}_{\text{avg}}(O)$

943 **Handling Uncertainty.** To maintain reliability and mitigate incorrect assumptions, the prompt ex-
944 plicitly instructs the model to use "unspecified" or "N/A" when an attribute value cannot be con-
945 fidently determined from available information. This approach preserves the integrity and trustwor-
946 thiness of the memory bank by avoiding low-confidence guesses.

947
948 **Compound Attributes and Consistency.** The prompt clearly addresses compound attributes (e.g.,
949 combining "fabric and wood" for material attributes), requiring these to be succinctly represented as
950 unified strings. Additionally, it underscores consistency across responses, ensuring that all outputs
951 adhere strictly to valid JSON formatting with appropriate string escaping. This structured consis-
952 tency facilitates seamless integration into the personalized memory bank infrastructure.

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

956
957
958 **A.1.2 ALGORITHM OF CALCULATING CROSS-USER OBJECT ATTRIBUTE SIMILARITY**
959

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

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

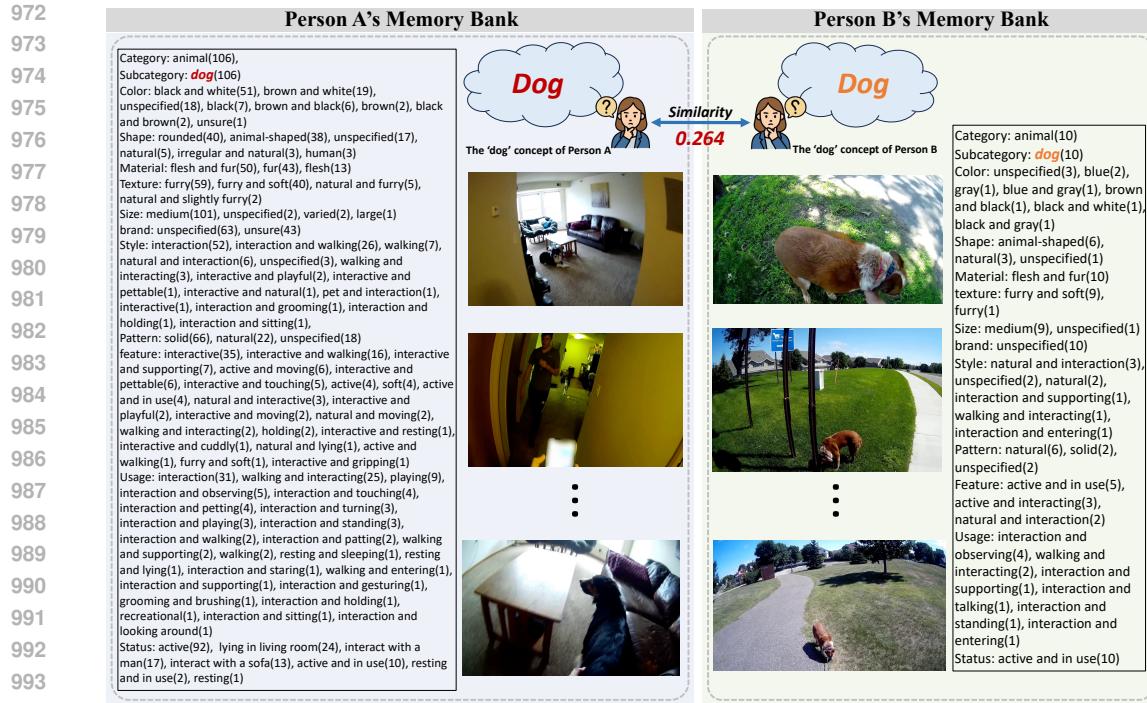


Figure 8: Qualitative comparison of the object concept “dog” across two users’ personal memory banks. Attribute distributions reveal substantial divergence in visual, contextual, and interactional properties. The computed similarity score (Jaccard index) is **0.264**, highlighting the necessity of modeling user-specific memory to capture personalized object semantics. .

A.1.3 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 8, 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.2 DETAILS FOR OUR EGORETRIEVER’S PROCESS

A.2.1 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

1026 **Algorithm 2** Computing Process of EgoRetriever for Personal Egocentric Video Retrieval

1027 **Input:** Natural language query Q , candidate clip set $\mathcal{C} = \{C_1, C_2, \dots, C_M\}$, user-specific semantic

1028 memory bank \mathcal{M} , reflective CoT prompt p_c .

1029 **Parameters:** Frozen Multimodal Large Language Model (MLLM) Ψ_M , frozen text encoder Ψ_T ,

1030 frozen video encoder Ψ_V .

1031 **Output:** Retrieved target clip C^* .

1032 1: Initialize pre-trained and frozen models Ψ_M, Ψ_T, Ψ_V .

1033 2: Retrieve personal context from memory bank: Query \mathcal{M} with Q to obtain personal object meta-

1034 data \mathcal{M}_q and reference image I_r .

1035 3: Generate target clip description using reflective CoT:

1036
$$T_t = \Psi_M(p_c \circ Q \circ \mathcal{M}_q \circ I_r)$$

1037

1038 4: Compute normalized text embedding for the target description:

1039
$$\hat{e}_T = \frac{\Psi_T(T_t)}{\|\Psi_T(T_t)\|_2}$$

1040

1041 5: Initialize a list of similarity scores $S = []$.

1042 6: **for** each candidate clip C_i in \mathcal{C} **do**

1043 7: Compute normalized video embedding for the candidate clip:

1044
$$\hat{e}_{V_i} = \frac{\Psi_V(C_i)}{\|\Psi_V(C_i)\|_2}$$

1045

1046 8: Compute similarity score: $s_i = \hat{e}_{V_i}^\top \hat{e}_T$ {Cosine similarity for normalized embeddings}

1047 9: Add s_i to S .

1048 10: **end for**

1049 11: Retrieve target clip: $C^* = \operatorname{argmax}_{C_i \in \mathcal{C}} S_i$ {Select clip with highest similarity score}

1050

1051 12: **return** C^*

1055

1056

1057 leverages the MLLM Ψ_M to consult the memory bank \mathcal{M} and retrieve pertinent personal object

1058 metadata \mathcal{M}_q and a reference frame I_r . Subsequently, using these contextual cues along with the

1059 original query Q and a reflective CoT prompt p_c , a detailed target clip description T_t is generated.

1060 This description T_t is then encoded using a text encoder Ψ_T . Each candidate clip $C_i \in \mathcal{C}$ is en-

1061 coded using a video encoder Ψ_V . The final retrieval of the target clip C^* is achieved by computing

1062 the cosine similarity between the encoded description and each encoded candidate clip. This ap-

1063 proach allows for a modular retrieval pipeline where the core reasoning and description generation

1064 are handled by the MLLM, independent of the specific video-language encoders used, requiring no

1065 additional training.

1066

1067

1068 A.2.2 COMPLETE TEMPLATE FOR REFLECTIVE CoT IN MULTIMODAL VIDEO RETRIEVAL

1069

1070 The complete template for our reflective Chain-of-Thought (CoT) reasoning prompt designed for

1071 multimodal video retrieval is detailed in Figure 9. This structured prompt systematically integrates

1072 visual observations, personalized object attributes, and user query intentions within a unified rea-

1073 soning framework. Initially, the *Original Image Description* step meticulously documents visual

1074 details from the provided reference frame, ensuring the inclusion of all relevant contextual cues.

1075 Subsequently, the *Thoughts* step explicitly interprets the user’s retrieval intention by analyzing the

1076 alignment of visual attributes with personalized object usage information. The subsequent *Reflec-*

1077 *tions* step involves a rigorous evaluation of identified visual and semantic elements, isolating those

1078 most congruent with the user’s implicit intent. Finally, the *Target Video Description* synthesizes

1079 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, pro-

moting coherent, efficient, and interpretable reasoning.

1080
 1081 **Original Image Description.** In this phase, the multimodal large language model (MLLM) is tasked
 1082 with comprehensively describing *all visible objects and their respective attributes* (e.g., *color, shape,*
 1083 *texture, size*). Additionally, the model must document *immediate surroundings and broader context-*
 1084 *ual factors (environmental conditions, indoor/outdoor setting)*, prioritizing precision and detail to
 1085 preserve critical visual evidence essential for subsequent analytical steps.

1086 **Thoughts.** Utilizing both the visual description and personalized object attributes (reflecting habitual
 1087 usage patterns), the MLLM explicitly *interprets the retrieval intent underlying the user’s query*.
 1088 It identifies and elaborates on visual elements (such as dominant colors, textures, or spatial config-
 1089 urations) closely aligning with the user’s specific object profile. Further, the MLLM incorporates
 1090 semantic considerations (such as temporal sequences or action relevance) essential to accurately
 1091 infer the retrieval context.

1092
 1093 **Reflections.** In this evaluative stage, the MLLM reexamines the highlighted visual and personal
 1094 object attributes from prior steps. The model critically *summarizes the integration of these visual*
 1095 *and usage details in informing its retrieval decision*. It explicitly highlights pivotal elements (e.g.,
 1096 distinguishing material characteristics, contextual environment) and articulates meta-reasoning jus-
 1097 tifications to ensure coherence between the reference imagery, object attributes summary, and the
 1098 user’s retrieval intention. It reflects precisely on the visual or usage-derived cues that underpin its
 1099 decision-making rationale.

1100
 1101 **Target Video Description.** Utilizing refined insights from the reflective analysis, the MLLM gener-
 1102 ates a concise and targeted description pinpointing the specific video segment containing the queried
 1103 object or interaction. This description is explicitly formulated as a *single, precise sentence* encom-
 1104 passing only retrieval-relevant elements, thus facilitating efficient and highly accurate retrieval.

1105 A.2.3 VISION-BY-LANGUAGE IN-CONTEXT LEARNING DETAILS

1106
 1107 Effectively executing Reflective Chain-of-Thought (CoT) reasoning in multimodal large language
 1108 models (MLLMs) requires not just general instructions but also concrete demonstrations of the rea-
 1109 soning process. To achieve this under a zero-shot setting without relying on direct visual guidance,
 1110 we adopt a vision-by-language in-context learning (ICL) strategy inspired by recent advances in
 1111 multimodal reasoning methodologies Wei et al. (2022); Mitra et al. (2024b); Zheng et al. (2023);
 1112 Tang et al. (2024a).

1113 Our Reflective CoT ICL provides MLLMs with structured language-based exemplars that guide
 1114 the model through each reasoning step solely via textual information. As depicted in Figure 12,
 1115 each example comprises clearly delineated components: an *Original Image Description*, *Thoughts*,
 1116 *Reflections*, and a *Target Video Description*.

1117 For clarity, consider the following example based on a user query and a provided object attributes
 1118 summary:

1119 **User Query:** *“Where was the dog after I laid the bed?”*

1120 **Visual Reference:** *Middle frame from a reference video that shows a bed’s large rectangular form*
 1121 *with a decorative patterned cover.*

1122 **Object Attributes Summary:** Detailed semantic attributes related to the bed, including aspects
 1123 such as “decorative patterned fabric,” “large size,” and “used and slightly messy” status.

1124 The Reflective CoT steps are as follows:

- 1125 • **Original Image Description:** The MLLM generates a detailed depiction of visually perti-
 1126 nent components to the user query. In this scenario, the description captures the bed’s large
 1127 rectangular form with a decorative patterned cover, noting its slightly messy state indicated
 1128 by creases and indentations, and contextualizes the scene within a residential bedroom with
 1129 daylight filtering through partially open curtains.
- 1130 • **Thoughts:** The model interprets the user’s intent, identifying the dog’s location post-
 1131 interaction with the bed. Leveraging details from the object attributes summary (e.g.,
 1132

1134
 1135
 1136
 1137 You are a highly skilled AI assistant specializing in multimodal video retrieval with deep chain-of-thought
 1138 reasoning. You will receive:
 1139
 1140 Visual Reference Image
 1141 – A single frame (the middle frame) from a candidate video.
 1142 – Contains key visual details: object color, shape, texture, size, and surrounding context (indoor/outdoor,
 1143 lighting, background scene).
 1144
 1145 Object Attributes Summary
 1146 – A concise personal profile of the object, including categorical and frequency data (e.g., major category,
 1147 subcategory, color combinations, material, style, usage frequency).
 1148 – Reflects the user's habitual usage and personal signature.
 1149
 1150 Your task is to produce a detailed chain-of-thought explanation and a final one-sentence description of the
 1151 target video. The target video is assumed to contain the object referenced in the user's query, based on both
 1152 the visual evidence from the candidate frame and the user's personal usage information.
 1153
 1154 Your response must be structured as a JSON object with the following keys:
 1155 {
 1156 "Original Image Description": <string>,
 1157 "Thoughts": <string>,
 1158 "Reflections": <string>,
 1159 "Target Video Description": <string>
 1160 }
 1161
 1162 **## Guidelines on Generating the Original Image Description**
 1163 - Provide a thorough and detailed description of the visual reference image.
 1164 - Describe all visible elements in the reference image: the object's attributes (color, shape, texture, size), its
 1165 immediate surroundings, and indoor/outdoor context.
 1166 - Be precise and comprehensive.
 1167
 1168 **## Guidelines on Generating the Thoughts**
 1169 - Explain your understanding of the user's query and the object attributes summary.
 1170 - Detail which visual cues (e.g., dominant colors, materials, spatial relations) align with the personal profile
 1171 - Consider semantic aspects such as Location/Positioning, Object Attributes, Temporal Sequence,
 1172 Presence/Absence and Action/Manipulation.
 1173 - Discuss which details in the candidate image were most influential in guiding your decision-making process.
 1174 - Conclude with how these insights were used to formulate your final target video description.
 1175
 1176 **## Guidelines on Generating the Reflections**
 1177 - Summarize how the integration of the visual clues and the object attributes influenced your approach.
 1178 - Highlight the most influential details (e.g., material, environment) and why they confirm the candidate video's
 1179 relevance.
 1180 - Explain how specific details (such as color, material, or setting) reinforced your decision.
 1181 - Concise meta-reasoning: justify key decisions that preserved coherence between the reference image, the
 1182 attribute summary, and the retrieval goal. Highlight which visual or personal-usage cues were decisive.
 1183 - Reflect on the overall impact of these considerations in crafting a logically connected and visually coherent
 1184 final description.
 1185
 1186 **## Guidelines on Generating the Target Video Description**
 1187 - Provide a single, concise sentence that identifies the most likely video segment containing the referenced
 1188 object.
 1189
 1190 Below is an example of the expected input and output formats:
 1191 ...
 1192
 1193

Figure 9: The complete template of our reflective Chain-of-Thought process for EgoMemory.

```

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197 Example Input:
1198 <Input>
1199 {
1200   "User Query": "Where was the dog after I laid the bed?"
1201   "Visual Reference": [Attached image showing the middle frame from a video,],
1202   "Object Attributes Summary":
1203     "subcategory: bed(48)"
1204     "color: varied and patterned(11), varied(8), unspecified(7), blue(4), multicolor and patterned(4), unsure(4), blue and
1205       multicolor(3), multicolor(3), beige and green(1), blue and white(1), beige and patterned(1), white and blue(1)"
1206     "shape: rectangular(33), rectangular and slightly padded(8), rectangular and cushioned(6), rectangular and slightly
1207       cushioned(1)"
1208     "material: fabric and wood(32), wood and fabric(9), fabric and synthetic(4), wood(2), wood and synthetic(1)"
1209     "texture: soft(11), soft and slightly patterned(10), smooth and slightly soft(7), soft and slightly wrinkled(6), soft and
1210       slightly smooth(3), smooth(3), soft and slightly textured(3), soft and rough(2), smooth and slightly worn(1), soft and
1211       smooth(1), quilted(1)"
1212     "size: large(45), medium(3)"
1213     "brand: unspecified(33), unsure(15)"
1214     "style: household and resting(31), residential(15), residential and resting(1), household and sleeping(1)"
1215     "pattern: decorative and patterned(19), solid(14), decorative(7), unspecified(4), varied(3), decorative and floral(1)"
1216     "feature: supporting and sleeping(18), resting(14), comfortable and supporting(10), supporting and resting(4),
1217       comfortable and supportive(1), adjustable(1)"
1218     "usage: resting and sleeping(19), sitting(12), resting and supporting(11), resting(2), interaction and placing(1), resting
1219       and lying(1), sleeping and resting(1), residential(1)"
1220     "status: used(15), used and in use(14), used and slightly untidy(9), used and slightly messy(7), unspecified(1), used and
1221       slightly wrinkled(1), used and slightly organized(1)"
1222   }
1223
1224 Example Output:
1225 <Response>
1226 {
1227   "Original Image Description": "A large rectangular bed with a decorative patterned fabric cover rests in a light-filled
1228     bedroom. The bed's surface is slightly messy, showing creases and indentations, with wooden bedframe legs visible. The
1229     surrounding room has residential furnishings, including a nightstand and partially open curtains allowing daylight.",
1230   "Thoughts": "The user's query focuses on the dog's location post-bed interaction. The Object Attributes Summary
1231     emphasizes the bed's 'used and slightly messy' status, 'decorative and patterned' design, and 'household and resting'
1232     context. The visual reference shows a residential bedroom setting with a disturbed bed surface, suggesting recent
1233     activity. The target video must encode both the bed's attributes (pattern, texture, usage state) and spatial context
1234     (bedroom) to align with retrieval model needs.",
1235   "Reflections": "The 'slightly messy' status from attributes and visible creases in the image strongly indicates recent bed
1236     interaction. The patterned fabric (mentioned in both color/texture attributes and visual description) provides distinctive
1237     visual anchors for retrieval models. Daylight through curtains in the reference image confirms the 'light-filled'
1238     environment. These elements collectively suggest the dog remained on the bed after interaction, with the patterned
1239     texture and spatial context being critical retrieval cues.",
1240   "Target Video Description": "A dog standing on a slightly messy, patterned bed in a light-filled bedroom."
1241 }

```

Figure 10: Examples of our vision-by-language in-context learning.

1242	Query	Reflective Chain-of-Thought	Retrieval Results
1243	Where are my gloves before I dropped them on the table?	<p>Thoughts: The intention is to focus on locating the gloves before they were dropped, which implies identifying their state and placement during use. The visual reference shows the gloves being worn and actively used, consistent with the 'protective' and 'used' attributes in the memory. The workshop context and upright posture suggest an ongoing task. Thus, the relevant clip must capture the gloves being worn just prior to their removal from the table.</p>	<p>A person wearing black gloves works on a task in a workshop before removing them and dropping them on a table.</p>
1244	Category: Clothing Subcategory: glove(17) Color: black (14), blue(3) Size: medium(17) Material: rubber(10), Usage: near a workshop table(10), ... Status: wearing and protecting hands(17), ...		
1245	Did I return my chair to its normal position?	<p>Thoughts: The intention is to confirm whether the gray chair was returned to its typical spot near the table. The personal memory highlights a gray chair with a C-shaped design. This suggests the chair had been moved and may later be repositioned</p>	<p>A C-shaped gray chair is moved from the middle of the room to near a table.</p>
1246	Category: Furniture Subcategory: chair(4) Color: gray (3), green(1) Shape: C-shaped (2), rectangular and slightly curved(1), Style: sitting(3), ... Usage: near a table (3), ... Status: in use(3), ...		
1247	What hammer did I collect from my colleague?	<p>Thoughts: The intention is to identify the specific hammer handed over by a colleague. The image reveals active collaboration in a construction setting, with one worker on a ladder and another assisting. Although the hammer itself is not clearly visible, the interaction, construction context, and proximity support the likelihood of a tool exchange. The personal memory highlights a 'blue and metallic' hammer typically used in striking and gripping, matching the presumed handoff action.</p>	<p>A blue and metallic hammer is handed over to a person wearing gloves in a construction setting.</p>
1248	Category: Tool Subcategory: hammer(16) Color: blue and metallic (3), ... Shape: T-shaped(8), ... Material: metal and wood(9), Usage: construction and interaction (12), ... Status: held and in use(10), ...		
1249	Who did I interact with when I was riding my bicycle?	<p>Thoughts: The intention is to identify the person the user interacted with while riding the bicycle, implying a focus on human interaction during active usage. The image confirms indoor parking and features the black bicycle among others. The target video should depict a direct interaction with another person, consistent with the interaction subcategory of usage, while the user is actively riding the black bicycle.</p>	<p>A person riding a black bicycle interacts with a person on an outdoor street.</p>
1250	Category: Bicycle Subcategory: bicycle(38) Color: black (15), ... Shape: elongated and mechanical(17), Usage: transportation and riding (27), riding in outdoor (5), ... Status: in use(36), stationary(2)		
1251			
1252			
1253			
1254			
1255			
1256			
1257			
1258			
1259			
1260			
1261			
1262			
1263			
1264			
1265			
1266			
1267			
1268			
1269			
1270			
1271			
1272			
1273			
1274			

Figure 11: Qualitative analysis demonstrating the advantages of reflective Chain-of-Thought (CoT) in interpreting user intent. Row 1 illustrates how reflective CoT focuses on the relevant context of gloves being worn and used before being dropped on a table, correctly inferring their usage despite potential visual clutter. Row 4 highlights how reflective CoT discards an incorrect assumption about indoor parking and instead focuses on identifying interactions during active bicycle usage outdoors. Reflective CoT enhances the accuracy of episodic memory retrieval by filtering out irrelevant details and aligning the reasoning process with the user’s true intent. Additional examples are provided in the sample data from our supplementary materials.

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.”

1296 This structured Reflective CoT approach enables the MLLM to systematically internalize reasoning patterns from textual exemplars, supporting consistent and accurate multimodal inference even
 1297 without direct visual references. By utilizing language-only in-context demonstrations, our method
 1298 effectively maintains training-free adaptability, enhancing retrieval accuracy through clearly articu-
 1299 lated reasoning pathways.
 1300

1302 A.2.4 MORE QUALITATIVE ANALYSIS OF REFLECTIVE CoT

1304 To demonstrate the advantages of reflective CoT in accurately interpreting user intent, we present
 1305 qualitative analyses in Figure 11. Reflective CoT plays a crucial role in filtering out irrelevant el-
 1306 ements, such as extraneous scene descriptions or hallucinated details, which are often distractions
 1307 in the reasoning process. This reflective process enhances the precision and reliability of episodic
 1308 memory retrieval. In Row 1, reflective CoT excels in focusing on the relevant context by identi-
 1309 fying the state and placement of gloves before being dropped on a table. Despite potential visual
 1310 clutter, reflective CoT correctly infers the gloves’ usage context, i.e., being worn and involved in
 1311 manual tasks, reinforcing the memory of the gloves as protective gear. Row 4 further illustrates the
 1312 power of reflective CoT by eliminating an incorrect assumption in the reasoning process. The image
 1313 initially suggests indoor parking and a black bicycle, but reflective CoT shifts focus to correctly
 1314 interpret the user’s query about interactions during active bicycle usage outdoors. This change in fo-
 1315 cus underscores the adaptability of reflective CoT in aligning with the user’s true intent, effectively
 1316 disregarding irrelevant contextual details.

1317 By integrating personal memory and contextual cues through explicit reflection, reflective CoT en-
 1318 sures a robust understanding of the user’s intent. This not only filters out extraneous elements but
 1319 also contributes to more accurate retrieval of relevant episodic memories, ultimately enhancing the
 1320 system’s interpretability and performance.

1321 A.3 DETAILS OF EGOMEMORY BENCHMARK CONSTRUCTION

1323 To rigorously evaluate memory-augmented *personalized* egocentric video retrieval, we introduce the
 1324 **EgoMemory** benchmark, constructed from Ego4D’s Natural Language Queries (NLQ) Grauman
 1325 et al. (2022). NLQ provides ~ 227 hours of head-mounted video from 137 participants across 74
 1326 locations, with free-form queries about “when/where/with whom/what,” closely reflecting realistic
 1327 recall scenarios.

1328 While NLQ offers rich content, its original design targets temporal localization within individ-
 1329 ual clips and does not aggregate multi-video, user-specific context. We therefore make personal-
 1330 ization *explicit*. Concretely, we operationalize personalization as *personally experienced objects*:
 1331 entities that (i) recur across a user’s videos and/or (ii) are linguistically tied to the user via first-
 1332 person/possessive/deictic cues. This definition does not require legal ownership; instead, it uses
 1333 recurrence and linguistic evidence to avoid conflating incidental scene exposure with genuine user
 1334 linkage.

1335 To address NLQ’s gaps, our benchmark treats each user as a distinct retrieval unit, aggregating all of
 1336 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)
 1337 **long-context participant selection**, requiring at least 10 videos and ≥ 1 hour cumulative footage
 1338 per user to ensure sufficient temporal breadth; (3) **manual verification**, where annotators review 20
 1339 additional clips (3s each) from the same user for the queried object class and label a query “personal”
 1340 if $\geq 90\%$ of reviewed instances match (“uncertain” if $\geq 75\%$). This yields 639 curated queries over 45
 1341 participants (245 videos), with $\sim 91.6\%$ labeled personal, thus emphasizing queries whose resolution
 1342 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
 1345 shown in the Figure) to keep queries whose main object is linguistically tied to the user (first-
 1346 person possessives or deictics), while allowing secondary impersonal objects. For example,
 1347 “What was the color of *my* drawstring bag?” is retained, whereas “In what aisle did I see a
 1348 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
 1349 and ≥ 1 hour cumulative duration. For retained queries, the referenced main object must recur

1350 across a user's videos, ensuring that answering the query draws on cross-video history rather than
 1351 a single clip.
 1352

3. **Manual verification.** For each query–target pair, annotators inspect 20 additional 3 s clips from
 1353 the same user containing the same object class. A query is “personal” if $\geq 90\%$ of reviewed
 1354 instances match the target object and “uncertain” if $\geq 75\%$. This process results in $\sim 91.6\%$
 1355 personal queries and reduces conflation with scene exposure.
 1356

1357

1358 You have samples of a query and the middle frame of the target video for an ****egocentric video**
 1359 **retrieval**** task, where each query may or may not represent a personalization query.
 1360 ****Personalization queries**** involve questions about objects or details closely related to the user's
 1361 personal belongings, personal actions, or personal spaces.
 1362 Analyze the given query ****step-by-step****, thinking carefully about whether it reflects a personal
 1363 connection or just a general interaction or observation.

1364 **## Guidelines on Generating the ****Thoughts******
 1365 1. ****Identify**** the object or event mentioned in the query.
 1366 2. ****Consider**** if the object/event typically belongs to, or is personally associated with, the user.
 1367 3. ****Determine**** if the query involves personal ownership, personal routine, or personal responsibility.
 1368 4. ****Distinguish**** between personal-related questions (about belongings, personal events, personal
 1369 spaces) and non-personal general observation questions.

1370 **## Output Instructions**
 1371 Based on your Thoughts, output your answer in ****JSON format****, clearly indicating **“YES”** if it is a
 1372 personalization query, or **“NO”** if it is not, along with your reasoning.

1373

1374 **## Input Format**
 1375 ````json`
 1376 `<Input>`
 1377 `{`
 1378 `“query”: “<User_Query>”,`
 1379 `“target_video”: “<Target_Video_Frame>”`
 1380 `}`
 1381 ``````

1382 **## Output Format**
 1383 ````json`
 1384 `<Response>`
 1385 `{`
 1386 `“personalization”: “YES” or “NO”,`
 1387 `“reason”: “<Your detailed reasoning>”`
 1388 `}`
 1389 ``````

1390 **## Example 1**
 1391 ````json`
 1392 `<Input>`
 1393 `{`
 1394 `“query”: “What was the color of my drawstring bag?”`
 1395 `}`
 1396 `<Response>`
 1397 `{`
 1398 `“personalization”: “YES”,`
 1399 `“reason”: “The query references a personal belonging (‘my drawstring bag’), clearly indicating personal`
 1400 `ownership.”`
 1401 `}`
 1402 ``````

Figure 12: The complete template of our CoT pre-screening.

1403 To address these gaps, our benchmark treats each user as a distinct retrieval unit, aggregating all of
 1404 their available videos as personal context at query time. We apply rigorous data filtering: participants

1404 must have multiple videos, and we use GPT-4o to select queries with long-context dependencies,
 1405 followed by manual curation (see Supplementary for details). The resulting dataset comprises 245
 1406 videos from 45 unique participants, spanning diverse everyday contexts and totaling 639 distinct
 1407 queries.

1408 Personalized memory banks are constructed for each participant as described in Section 3.2, resulting
 1409 in 165,795 user-specific object annotations. The number of unique object types per participant
 1410 ranges from 59 to 638 (median: 129), with memory bank sizes ranging from 322 to 10,454 annotations
 1411 (median: 1,312). Figure 3 visualizes the most frequent objects.

1412 The candidate retrieval set includes 2,228 video clips extracted from participants’ historical ego-
 1413 centric videos. Each user’s average candidate set size is approximately 33 clips, capturing a rich
 1414 spectrum of personal and habitual contexts. Video lengths within EgoMemory span from as short
 1415 as 4 seconds to a maximum of 300 seconds, averaging 103.82 seconds per clip. The total cumulative
 1416 duration of the candidate set amounts to 64.25 hours. Notably, these varied temporal spans
 1417 enable thorough assessments of retrieval robustness across different episodic memory lengths and
 1418 complexities.

1420 A.4 EXPERIMENT DETAILS ON EGO-CVR BENCHMARK

1422 A.4.1 EGO-CVR BENCHMARK OVERVIEW

1424 To further assess the generalization ability of our proposed EgoRetriever framework, we con-
 1425 duct additional experiments on the EgoCVR benchmark Hummel et al. (2024). EgoCVR is an ego-
 1426 centric, composed video retrieval dataset designed to evaluate a model’s capability to generate tar-
 1427 get video descriptions conditioned on reference images and textual modifications. This benchmark
 1428 presents unique challenges by emphasizing nuanced temporal and semantic variations in first-person
 1429 video data, complementing our primary evaluations on EgoMemory.

1430 A.4.2 TASK DEFINITION.

1432 In EgoCVR, the Composed Video Retrieval (CVR) task is defined as follows: Given a query com-
 1433 prising a reference frame (sampled from a video) and a free-form textual modification instruction,
 1434 the objective is to retrieve the corresponding target video clip from a gallery. This retrieval is per-
 1435 formed under two conditions: a *global* setting, where the gallery includes all test clips, and a *local*
 1436 setting, where the gallery is restricted to clips from the same long-form video as the query. Formally,
 1437 let \mathcal{I} denote the set of reference images, \mathcal{T} the set of textual modifications, and \mathcal{V} the set of candi-
 1438 date videos. For each query (i_q, t_q) , the task is to identify $v^* \in \mathcal{V}$ that best reflects the semantic
 1439 transformation specified by t_q when applied to i_q .

1440 A.4.3 DATASET CONSTRUCTION.

1442 The EgoCVR dataset consists of egocentric videos capturing a diverse array of daily activities and
 1443 object manipulations. Each annotated instance comprises:

- 1445 • A **reference image**: a single frame extracted from the query video.
- 1446 • A **textual modification**: a concise natural language instruction describing the intended
 1447 change (e.g., “use a different object,” “perform the action faster”).
- 1448 • A **target video**: a short clip (2–8 seconds) from the dataset that realizes the specified
 1449 modification.

1451 Each query is paired with ground-truth target(s), and standard test splits are provided to ensure com-
 1452 parability across methods. For further details on dataset construction, we refer the reader to Hummel
 1453 et al. (2024).

1454 A.4.4 EVALUATION PROTOCOL

1455 We follow the official evaluation protocol proposed in Hummel et al. (2024). Specifically, two
 1456 retrieval settings are considered:

1458 Table 4: Analysis the generalization of our EgoRetriever on the EgoCVR benchmark.
1459

Method	Global			Local		
	R@1	R@5	R@10	R@1	R@2	R@3
CIR-VL Karthik et al. (2024)	2.0	6.8	10.6	33.6	49.7	61.4
OSrCIR Tang et al. (2024a)	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

- **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.

1472 For each query, the model generates a target description based on the reference image and textual
1473 modification, then retrieves the most relevant clip from the gallery.

1474 A.4.5 EVALUATION METRICS

1476 Performance is measured using standard top- k retrieval metrics:

- **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.

1481 Following the official setting Hummel et al. (2024), we report R@1, R@5, and R@10 for the global
1482 setting, and R@1, R@2, and R@3 for the local setting, following the original benchmark protocol.
1483 Metrics are averaged across all test queries.

1485 A.4.6 EXPERIMENTAL SETUP

1487 All experiments are conducted on the official EgoCVR test splits, strictly adhering to the standardized
1488 evaluation settings. Model implementations and hyperparameters follow those described in
1489 Section of the main paper. The primary methods compared include:

- **CIR-VL** Karthik et al. (2024): A training-free composed image retrieval approach adapted for video.
- **OSrCIR** Tang et al. (2024a): A one-stage, training-free composed retrieval framework.
- **TFR-CVR** Hummel et al. (2024): A two-stage approach utilizing video captioning and LLM-based modification.
- **TFR-CVR***: A variant using GPT-4o as both captioner and modifier.
- **EgoRetriever**: Our proposed one-stage, training-free framework utilizing multimodal large language models (MLLMs) and reflective chain-of-thought reasoning for precise target description generation.

1501 A.4.7 RESULTS

1503 As summarized in Table 4 of the main paper, EgoRetriever achieves state-of-the-art performance
1504 on EgoCVR across both global and local retrieval settings. Our method consistently outperforms
1505 all baselines in Recall@K, highlighting its superior ability to generalize to novel egocentric
1506 video retrieval tasks.

1507 A.4.8 COMPARISON OF INPUT MODALITIES: OUR BENCHMARK VS. EGOCVR

1509 While both our proposed benchmark (EgoMemory) and the EgoCVR benchmark Hummel et al.
1510 (2024) focus on the challenging task of composed video retrieval, a key distinction lies in the design
1511 of their input modalities and the manner in which user intent and contextual information are encoded
for retrieval.

1512 **Our Benchmark (EgoMemory):** Our evaluation protocol is motivated by real-world episodic
 1513 memory retrieval, wherein the user’s search intent is inherently personal and context-dependent. To
 1514 capture this, we design queries to include three components:
 1515

- 1516 • **Textual Memory Bank:** A collection of personal, user-centric text snippets that serve as
 1517 a long-term memory repository, reflecting frequently encountered objects, past actions, or
 1518 unique user experiences. This memory bank enables models to retrieve and reason over
 1519 personalized historical context during retrieval.
- 1520 • **Reference Image:** A visual snapshot or frame representing the starting point of the query,
 1521 grounding the search in a specific visual context.
- 1522 • **User Query:** A free-form natural language request, which may include references to per-
 1523 sonal context, intent, or temporal cues (e.g., “Find when I last used the red mug in the
 1524 kitchen”).

1525 This multimodal input setting allows models to perform retrieval that is both visually grounded and
 1526 deeply personalized, supporting rich reasoning over long-context egocentric video archives.
 1527

1528 **EgoCVR Benchmark.** By contrast, the EgoCVR benchmark Hummel et al. (2024) employs a
 1529 more constrained input format, where each query consists of:
 1530

- 1531 • **Reference Image:** A single frame from the query video, serving as the visual anchor for
 1532 retrieval.
- 1533 • **Modification Text:** A concise natural language instruction specifying a semantic change
 1534 or action to be performed (e.g., “pick up a different object,” “use the other hand”).

1535 Notably, EgoCVR does not provide explicit access to personalized memory or long-term user con-
 1536 text; all retrievals are based solely on the visual and textual modifications present in the immediate
 1537 query. This design tests the model’s ability to interpret and execute fine-grained semantic transfor-
 1538 mations but does not directly address personalization or long-horizon reasoning.
 1539

1540 **Summary of Differences.** The primary distinction is that EgoMemory introduces a *textual mem-
 1541 ory bank* and *user query* to explicitly model personal context and long-term memory, supporting
 1542 retrieval scenarios that are more representative of real-world egocentric memory augmentation. In
 1543 contrast, EgoCVR focuses on visually-anchored modifications without leveraging historical or per-
 1544 sonalized information. Our experimental evaluation on both benchmarks demonstrates the gener-
 1545 alization of our approach to settings with and without access to external memory, highlighting the
 1546 flexibility and robustness of our retrieval framework.
 1547

1548 A.5 EXPERIMENT DETAILS ON EGOLIFEQA BENCHMARK

1549 A.5.1 EGOLIFEQA BENCHMARK OVERVIEW

1551 EgoLife is a week-long, multi-person egocentric study in which six participants cohabit and con-
 1552 tinuously record with AI glasses, producing a comprehensive ~300-hour multimodal dataset; Ego-
 1553 LifeQA is a derived QA suite tailored to long-context assistance. The released EgoLifeQA subset
 1554 contains 6,000 questions over 266 hours of video.
 1555

1556 A.5.2 TASK DEFINITION

1557 EgoLifeQA defines five question types: *EntityLog* (objects and their attributes/locations), *Even-
 1558 tRecall* (past event details), *HabitInsight* (personal habit patterns), *RelationMap* (interpersonal in-
 1559 teractions), and *TaskMaster* (task tracking and reminders). Each item is multiple-choice and is
 1560 constructed to require evidence from at least five minutes prior to the question time.
 1561

1562 A.5.3 DATASET CONSTRUCTION

1563 Long, “visual–audio” captions are first generated and fed to GPT-4o to propose timestamped QA
 1564 candidates; annotators then filter and refine them, retaining only questions with long-range depen-
 1565 dencies (at least five minutes) and high real-world relevance, yielding the final QA set.
 1566

1566
1567

A.5.4 EVALUATION PROTOCOL

1568
1569

We follow the official setting: models must answer the fixed-choice questions using only the benchmark inputs and produce supporting predictions per category. :contentReference[oaicite:7]index=7

1570
1571

A.5.5 EVALUATION METRICS

1572
1573
1574

Performance is reported as accuracy (%) per question type, averaged over the official test questions (EntityLog, EventRecall, HabitInsight, plus the remaining categories as applicable).

1575
1576

A.5.6 EXPERIMENTAL SETUP

1577
1578
1579
1580

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. :contentReference[oaicite:9]index=9

1581
1582

A.5.7 RESULTS

1583
1584
1585
1586

As summarized in Table 4 EgoRetriever attains the highest accuracy on **EntityLog**, **EventRecall**, and **HabitInsight** (42.5, 45.1, 37.7), outperforming the strongest baseline by +3.3, +3.0, and +6.6 points, respectively, demonstrating transfer beyond retrieval to fixed-answer, long-context QA focused on entities, events, and habits.

1587
1588
1589

A.6 MORE ABLATION STUDY

1590
1591

Table 5: 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.

1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

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 5 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-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’**

1620
1621 Table 6: Comparative analysis of computational cost, latency, memory usage, API expenditure, and
1622 retrieval performance across baseline and proposed models.
1623

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

1629
1630 **evaluate alternative reference frame selection strategies.** Compared to the randomly selected
1631 frame baseline (model ‘7’), our middle-frame selection strategy (model ‘1’) provides a slightly im-
1632 proved retrieval performance (0.427% higher on average). An alternative frame selection method,
1633 which leverages clip embeddings to calculate the similarity between captions and individual frames
1634 (model ‘8’), achieves marginally higher performance (0.313% higher on average). However, this
1635 embedding-based approach significantly reduces efficiency due to the computational overhead of
1636 evaluating similarity against dense frame-level descriptions provided by Ego4D’s narrations (at 30
1637 fps). Therefore, our simpler middle-frame strategy is optimal, striking an effective balance between
1638 retrieval accuracy and computational efficiency. **(4) Model ‘9-10’ analysis the dependence on hu-
1639 man narrations of our personal memory bank.** Dropping 25% (model ‘9’) and 50% (model ‘10’)
1640 of human narrations yields moderate degradations, consistent with redundancy in our memory bank:
1641 recurring personal objects are recorded across clips and frequency-based filtering suppresses noise
1642 from missing narrations.

1643
1644 A.7 DETAILED COMPARISON OF COMPUTATIONAL COST

1645 In Table 6 we report a four-way cost-effectiveness analysis, mean query latency (averaged over 100
1646 runs), peak GPU memory at inference, OpenAI API expenditure per call, and average retrieval ac-
1647 curacy (ViT-L/14 backbone) on EgoMemory and EgoCVR, covering two prior baselines (CIReVL
1648 and OSrCIR), the two-stage TFR-CVR, and our single-stage EgoRetriever variants. CIReVL,
1649 which relies on a GPT-3.5 captioning head plus a separate retrieval LLM, incurs the highest mem-
1650 ory footprint (40 GB) and the slowest response time (~1.4 s), despite its modest API fee, and
1651 delivers the lowest accuracy (24.86). OSrCIR and TFR-CVR both reduce memory to 16 GB by
1652 avoiding a BLIP-2 captioner; however, OSrCIR’s one-stage design halves latency to ~0.7 s but
1653 still trails TFR-CVR by 13.4 mR points. Our EgoRetriever with GPT-4o-mini preserves OSr-
1654 CIR’s low memory/latency profile while lifting performance to 46.92, and the full GPT-4o variant
1655 maintains this latency (~ 0.7 ± 0.08s) while achieving the highest accuracy (48.19) without in-
1656 creasing memory consumption. Taken together, these results show that (i) eliminating separate
1657 captioning modules yields substantial memory and speed benefits, (ii) mini-scale MLLMs already
1658 strike an excellent cost-performance balance, and (iii) our reflective one-stage pipeline offers the
1659 best absolute accuracy with competitive operational costs, underscoring its suitability for real-time,
1660 resource-constrained egocentric retrieval systems.

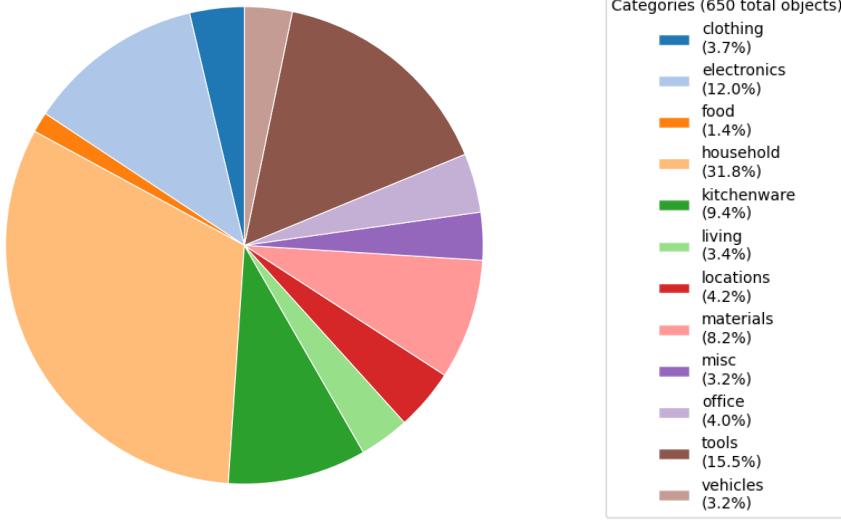
1661 A.8 SCALABILITY UNDER INCREASING VIDEO/METADATA VOLUME
16621663 Table 7: Scalability with dataset size. Latency is averaged over 100 queries; API cost is per call.
1664

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

1670 **Findings.** As Table 7 shows, query-time latency remains ≈0.65–0.70 s with small variance as the
1671 corpus grows from 10 to 245 videos; peak GPU memory is flat at 16 GB and per-query API ex-
1672 penditure is unchanged. This stability stems from EgoRetriever using precomputed video and
1673 memory embeddings, so online retrieval reduces to lightweight lookups and similarity scoring; only

1674 offline indexing scales with data volume and does not affect interactive latency. In tandem with
 1675 Fig. 6 (main), where larger memory-bank context improves accuracy, these results indicate that
 1676 EgoRetriever scales efficiently in both cost and effectiveness as personalized data accumulates.
 1677

1679 A.9 CANDIDATE SET STATISTICS



1699 Figure 13: Distribution of 650 object instances across 13 semantic categories in the candidate re-
 1700 trieval set. The largest categories are household (31.8%), tools (15.5%), and electronics (12.0%),
 1701 reflecting the diversity and realism of egocentric video environments.

1702
 1703 To facilitate personalized episodic memory retrieval, we construct a candidate retrieval set for each
 1704 participant by aggregating their historical egocentric video clips prior to each query. This design
 1705 simulates realistic personal memory banks and enables a thorough evaluation of retrieval models
 1706 across diverse user contexts.
 1707

1708 **Composition and Scale.** The benchmark candidate set comprises a total of 2,228 video clips col-
 1709 lected from 45 unique users, sampled from their continuous head-mounted video recordings. The
 1710 number of candidate clips per participant ranges from 9 to 61, with an average of 33 clips per user.
 1711 Clip durations vary from 4 to 300 seconds, with a mean duration of 103.82 seconds, resulting in a
 1712 cumulative total of 64.25 hours of video data in the candidate pool.

1713 **Object Instance Distribution.** Within the candidate set, we annotated 650 object instances span-
 1714 ning 13 semantic categories. Figure 13 illustrates the distribution of these object categories. The
 1715 largest proportions are household objects (31.8%), tools (15.5%), and electronics (12.0%), alongside
 1716 kitchenware (9.4%), materials (8.2%), and other everyday object classes. This distribution reflects
 1717 the diversity and complexity of real-world personal environments encountered in egocentric video.

1718 **Personalization and Diversity.** To assess diversity, we computed Jaccard similarity scores for the
 1719 top 100 most frequent object types based on their attribute metadata. Inter-participant similarity
 1720 was below 0.4 for 74% of object types, indicating high heterogeneity between users. Furthermore,
 1721 intra-participant similarity was consistently higher than inter-participant similarity, underscoring the
 1722 personalized and user-specific nature of each candidate set.

1723 **Evaluation Metric.** For all experiments, we report **mean Recall@K** ($K=1, 2, 3$), computed sepa-
 1724 rately for each user over their candidate set and then macro-averaged across users. This approach,
 1725 consistent with prior benchmarks Grauman et al. (2022); Hummel et al. (2024), ensures that per-
 1726 formance reflects the retrieval difficulty for each individual and mitigates biases due to uneven query
 1727 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.

1728 A.10 MORE IMPLEMENTATION DETAILS
17291730 A.10.1 DETAILS OF RIGOROUS DATA FILTERING
1731

1732 To ensure rigorous evaluation and the authenticity of personalized retrieval contexts in the **EgoMem-
1733 ory** benchmark, we implemented a comprehensive data filtering process. Initially, we screened the
1734 Ego4D Natural Language Queries (NLQ) dataset Grauman et al. (2022), restricting inclusion criteria
1735 to participants possessing multiple egocentric videos. This criterion was essential to authentically
1736 simulate realistic retrieval scenarios, as genuine personal retrieval contexts inherently involve mul-
1737 tiple video interactions over extended periods.

1738 Next, we utilized GPT-4o to systematically identify and select queries exhibiting clear long-context
1739 dependencies. Specifically, GPT-4o was prompted to assess each query’s dependency on temporal
1740 context beyond immediate video boundaries, prioritizing queries whose interpretations or resolu-
1741 tions necessitate referencing historical, user-specific contexts. Examples of selected queries typi-
1742 cally involved repeated interactions, habitual activities, or persistent object engagements spanning
1743 multiple video sessions.

1744 Following the AI-driven initial screening, we conducted meticulous manual curation to verify query
1745 suitability rigorously. Expert annotators reviewed each GPT-4o-identified query, confirming gen-
1746 uine long-context relevance and filtering out queries ambiguous in contextual dependence or insuf-
1747 ficiently representative of personalized retrieval demands. Critically, filtered samples required that
1748 participants have various candidate video clips, specifically more than 10 videos per participant,
1749 with an average candidate set size of approximately 33 clips. This ensured a robust and diverse
1750 long-context retrieval environment, capturing rich habitual and episodic nuances.

1751 The resulting refined dataset comprises 245 rigorously filtered videos from 45 participants, provid-
1752 ing 639 distinct, carefully validated long-context queries. This comprehensive filtering approach
1753 significantly enhances the benchmark’s effectiveness in accurately assessing personalized retrieval
1754 capabilities under authentic, user-specific episodic memory contexts.

1755 A.10.2 EXTENDED IMPLEMENTATION DETAILS FOR EGOMEMORY AND EGORETRIEVER
1756

1757 We leveraged GPT-4o extensively to construct detailed user-specific memory banks by generating
1758 comprehensive object-centric metadata annotations from video clips. Additionally, GPT-4o facili-
1759 tated reflective Chain-of-Thought (CoT) reasoning integral to our retrieval framework. Specifically,
1760 we conducted these annotation and reasoning processes for 45 participants, covering a total of 245
1761 videos, over approximately one month. This substantial annotation effort underscores both the com-
1762 putational and temporal investments involved in establishing robust personalized memory contexts.

1763 Our retrieval experiments utilized four NVIDIA V100 GPUs, each equipped with 32GB memory.
1764 We evaluated multiple state-of-the-art video-language models, including LanguageBind Zhu et al.
1765 (2023), CLIP Radford et al. (2021), BLIP Li et al. (2022), and EgoVLPv2 Pramanick et al. (2023).
1766 Within our proposed EgoRetriever, EgoVLPv2 served as the primary text encoder. To represent
1767 videos visually, embeddings from CLIP and BLIP were computed by averaging across embeddings
1768 extracted from 15 uniformly sampled frames per video clip. For each candidate video, these vi-
1769 sual embeddings were then matched with the GPT-4o-generated textual descriptions using cosine
1770 similarity to determine retrieval accuracy. For further detailed methodology and specific parameter
1771 settings, please refer to the supplementary materials.

1772 A.11 LIMITATIONS
1773

1774 Although EgoRetriever demonstrates strong performance and significant advances in person-
1775 alized long-context egocentric video retrieval, several limitations warrant consideration. First, our
1776 model critically depends on accurate extraction and annotation of user-specific object-centric meta-
1777 data via Multimodal Large Language Models (MLLMs), implicitly assuming these annotations to
1778 be precise and exhaustive; inaccuracies or omissions could substantially degrade retrieval effec-
1779 tiveness. Second, as highlighted by our failure analysis, the framework frequently encounters dif-
1780 ficulties in object disambiguation within cluttered scenes and struggles with context interpretation
1781 when visual cues are ambiguous or insufficient, indicating sensitivity to visual clarity and query
specificity. Third, the Ego4D dataset, from which our benchmarks are constructed, lacks explicit

1782 temporal annotations across different videos for individual participants, limiting the capability to
 1783 accurately reconstruct chronological timelines of user experiences, which is essential for genuinely
 1784 long-context episodic retrieval. Furthermore, while our evaluation validates generalizability using
 1785 the EgoMemory and EgoCVR benchmarks derived from Ego4D, the method has not yet been exten-
 1786 sively assessed across other diverse egocentric datasets, potentially constraining broader applicabil-
 1787 ity. Finally, the reliance on computationally intensive MLLMs and extensive metadata annotations
 1788 raises scalability and efficiency concerns, particularly for deployment on resource-limited wearable
 1789 devices, emphasizing the necessity of future research into streamlined model optimization and an-
 1790 notation methodologies.

1791 A.12 BROADER IMPACTS

1792 Our proposed framework, *EgoRetriever*, introduces significant advancements toward personal-
 1793 alized, long-context episodic memory retrieval, opening promising avenues for augmented human
 1794 cognition and improved assistive technologies. However, its deployment also brings potential so-
 1795 cietal risks warranting careful consideration. First, extensive recording and storage of personalized
 1796 egocentric data inherently pose substantial privacy concerns, as such continuous visual and context-
 1797 ual capture could inadvertently reveal sensitive personal information or be exploited for unautho-
 1798 rized surveillance and tracking. Secondly, even when functioning correctly, the model might unin-
 1799 tentionally reinforce biases embedded in training data or annotations, possibly leading to unequal
 1800 performance across diverse demographic groups, thus raising fairness considerations. Additionally,
 1801 misuse of the proposed memory-augmented retrieval technology could facilitate invasive monitoring
 1802 or targeted manipulation based on personal habits and behaviors, resulting in malicious outcomes
 1803 such as stalking, identity theft, or psychological manipulation. Incorrect retrieval outcomes, particu-
 1804 larly involving sensitive contexts or critical decisions, could also lead to harmful personal or societal
 1805 consequences. To mitigate these risks, we advocate implementing robust privacy-preserving mea-
 1806 sures, including data encryption, strict access controls, and user-centric data ownership frameworks.
 1807 Moreover, comprehensive fairness audits and continuous model evaluations across diverse user pop-
 1808 ulations are essential to ensure equitable deployment. Finally, careful consideration of transparent
 1809 usage policies and developing detection mechanisms for identifying and preventing misuse are cru-
 1810 cial steps toward responsibly harnessing the full potential of personalized egocentric video retrieval
 1811 technologies.

1812 A.13 EXTENDED RELATED WORKS

1813 **Multimodal Chain-of-Thought Reasoning for Egocentric Video Retrieval.** Multimodal Large
 1814 Language Models (MLLMs) have recently demonstrated remarkable reasoning capabilities, partic-
 1815 ularly when equipped with Chain-of-Thought (CoT) prompting strategies Wei et al. (2022); Kojima
 1816 et al. (2022); Zheng et al. (2023); Zhang et al. (2023). In the domain of composed image retrieval
 1817 (CIR), Recent advances introduced OSrCIR Tang et al. (2024a), a training-free, one-stage reflective
 1818 CoT framework that enables MLLMs to reason about manipulation intent and preserve contextual
 1819 information, thereby improving retrieval accuracy without the typical information loss associated
 1820 with two-stage approaches. Similarly, EmbodiedGPT Mu et al. (2023) extends CoT prompting to
 1821 vision-language pre-training in embodied environments, leveraging an EgoCOT dataset and prefix-
 1822 tuned LLMs to enable agents to perform complex action planning through multimodal reasoning.
 1823 Despite these advances, existing research on episodic memory retrieval from egocentric videos has
 1824 predominantly focused on short-term or single-video scenarios Grauman et al. (2022); Hummel
 1825 et al. (2024); Yang et al. (2025), with limited attention to the long-context, personalized nature in-
 1826 trinsic to human memory. Current benchmarks lack comprehensive personal memory banks and
 1827 rich, user-specific object annotations Singh et al. (2016); Damen et al. (2020); Núñez-Marcos et al.
 1828 (2022); Wang et al. (2023b), which are critical for modeling real-world episodic recall. Building
 1829 upon recent advances in multimodal CoT reasoning Tang et al. (2024a); Mu et al. (2023); Mitra
 1830 et al. (2024a); Zheng et al. (2023); Zhang et al. (2024), our work is the first to bring this paradigm
 1831 to personalized egocentric video retrieval. We introduce *EgoRetriever*, a training-free frame-
 1832 work that combines MLLMs with reflective CoT prompting, leveraging a comprehensive personal
 1833 memory bank constructed from extensive user-specific object annotations in Ego4D Grauman et al.
 1834 (2022). This enables the model to interpret nuanced user queries and generate detailed, contextually
 1835 grounded descriptions of target video clips by incorporating recurring personal objects, habitual ac-
 1836 tivities, and social interactions. Extensive experiments on the EgoMemory and EgoCVR Hummel

1836 et al. (2024) benchmarks demonstrate that `EgoRetriever` consistently outperforms existing base-
 1837 lines, underlining its strong generalizability and promise for real-world deployment in personalized
 1838 episodic memory retrieval.
 1839

1840 **A.14 THE USE OF LARGE LANGUAGE MODELS (LLMs)**
 1841

1842 LLMs were used in two distinct ways: as core research components of our method and as gen-
 1843 eral-purpose writing aids. The authors take full responsibility for all content; LLMs are not authors.
 1844

1845 **Role in research methodology.** We use multimodal LLMs strictly at inference time, without
 1846 fine-tuning, as part of our proposed system: (i) GPT-4o and GPT-4o-mini are used for reflective
 1847 chain-of-thought reasoning to generate target video descriptions, (ii) GPT-4o is used for CoT pre-
 1848 screening of NLQ queries during dataset filtering, and (iii) LLaVA and Qwen2.5-VL serve as open-
 1849 source alternatives in ablations. Prompts, pipelines, and hyperparameters are provided in the Ap-
 1850 pendix for reproducibility. When hosted APIs were used, we transmitted only minimal, structured
 1851 metadata (and reference frames where explicitly stated), never raw egocentric streams.
 1852

1853 **Role in writing.** LLMs (GPT-4 class) were used to assist with copy-editing, grammar, and word-
 1854 ing suggestions (e.g., tightening paragraphs, standardizing terminology, improving figure/table cap-
 1855 tions). They did not originate scientific claims, experimental designs, or citations; all technical
 1856 content, analyses, and references were authored, verified, and curated by the authors.
 1857

1858 **Accountability and authorship.** All results, claims, and text are the responsibility of the authors.
 1859 We checked model outputs for accuracy and potential plagiarism. LLMs are not eligible for author-
 1860 ship and are credited here solely as tools.
 1861

1862 **Data handling.** We avoided sharing personally identifying content with hosted models; no user
 1863 identities, faces, audio, GPS traces, or full-length videos were transmitted. Open-source MLLMs
 1864 enable fully local execution when stricter privacy is required.
 1865

1866 **Reproducibility.** We document model versions, prompts, and settings (Appendix) and note the
 1867 inherent non-determinism of API LLMs. Any deviations are reported in the experimental details.
 1868
 1869
 1870
 1871
 1872
 1873
 1874
 1875
 1876
 1877
 1878
 1879
 1880
 1881
 1882
 1883
 1884
 1885
 1886
 1887
 1888
 1889