Vinci2: Providing Proactive Assistance in Continuous Egocentric Videos

Sitong Gong[1,2,3], Tianyu Yan[⋆1,2,3], Caixin Kang[⋆2,3], Bo Zheng[2], Xiang Ruan[1], Huchuan Lu[1], Kaipeng Zhang[2], Yoichi Sato[3], and Yifei Huang[†2,3]

[1]Dalian University of Technology [2]Alaya Lab [3]The University of Tokyo
stgong@mail.dlut.edu.cn, {tianyu.yan, caixin.kang, bo.zheng, kaipeng.zhang}@shanda.com, {ruanxiang, lhchuan}@dlut.edu.cn, ysato@iis.u-tokyo.ac.jp, hyf015@gmail.com

Abstract. When should an intelligent assistant speak up without being asked? Continuous egocentric video offers rich, evolving context that enables a new form of assistance: one that is proactive rather than merely reactive. Yet existing approaches either wait passively for user queries or treat every detected event as requiring a response, without considering the user's history, current activity, or whether assistance would actually be welcome. We reframe proactive assistance as a context-dependent decision problem: the agent must not only perceive what is happening, but reason over accumulated temporal context to determine when and whether to intervene. To this end, we present Vinci2, a proactive egocentric assistance system that advances the on-device assistant Vinci from reactive response toward proactivity. On the evaluation side, we present EgoServe, the first large-scale benchmark for proactive assistance in continuous egocentric video. EgoServe comprises over 3,000 service instances organized along 4 temporal memory horizons, ranging from immediate safety alerts to long-term habit coaching, across 10 service categories. On the modeling side, we propose EgoMemo, a training-free, memory-augmented agent that maintains three complementary memory representations: multi-scale temporal summaries, a semantic knowledge graph, and visual embedding archives. At each timestep, EgoMemo performs retrieval-augmented reasoning to determine whether assistance is warranted and, if so, produces contextually grounded responses. Experiments demonstrate that EgoMemo establishes strong baselines on EgoServe while remaining competitive on existing egocentric benchmarks. Our benchmark and code are publicly available at Vinci2.

Keywords: Egocentric Vision · Proactive VLM · LLM Agent

⋆ Equal contribution.
† Corresponding author.

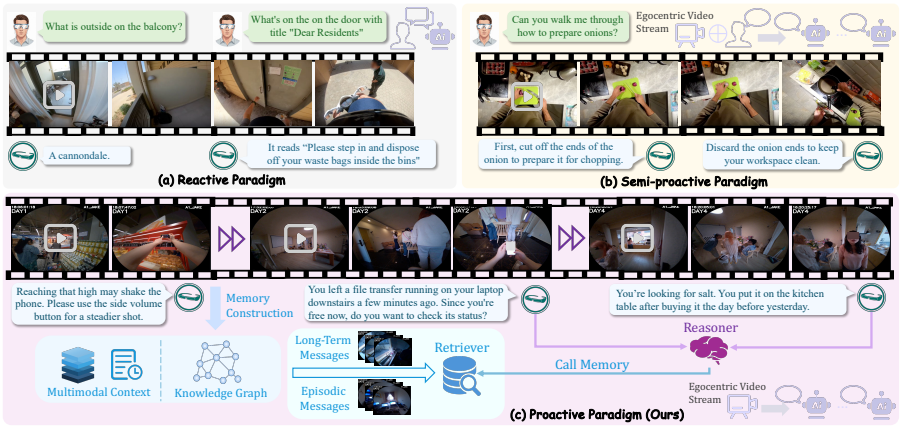

Fig. 1: Three paradigms of egocentric assistants. (a) Reactive: responds only to explicit user queries. (b) Semi-proactive: monitors the video stream for predefined task-relevant events. (c) Proactive (ours): autonomously decides when and how to intervene without any user prompt.

1 Introduction

The promise of proactive egocentric intelligence is an assistant that sees what you see, understands your context as it evolves, and offers help at the right moment without being asked. Recent progress in Video-LLMs [28, 31, 48, 50, 51], streaming visual perception [25, 43, 65], and egocentric foundation models [20, 59, 64] has brought this vision within reach, enabling continuous comprehension and reasoning over first-person video. Yet while the perception capabilities are maturing rapidly, the question of how and when an assistant should proactively intervene remains largely unaddressed.

Existing egocentric assistants operate under two limiting paradigms, as illustrated in Fig. 1. Most current Video-LLMs [28, 44, 71] follow a *reactive* paradigm, responding only when explicitly prompted. Recent event-triggered systems [54, 68, 69] adopt a *semi-proactive* paradigm: given a task instruction provided by the user in advance, they monitor the video stream for predefined events and generate responses upon detection. However, these methods are constrained by the scope of the initial instruction and the immediate visual context, lacking the ability to reason over long-horizon historical observations, and offering no mechanism to assess whether the current situation genuinely warrants interrupting the user. We argue for a third paradigm, *proactive* assistance, that has not yet been explored: the agent reasons over the user's accumulated context, including their history, habits, current activity, and goals, to make a deliberate decision about whether, when, and how to intervene. We instantiate this paradigm in Vinci2, a successor to the egocentric assistant Vinci [20] that advances from reactive response to genuine proactivity. Vinci2 comprises two complementary components: a benchmark **EgoServe**, and a training-free agent **EgoMemo**, which we detail below.

On the evaluation side, no existing benchmark addresses this need. Offline video QA benchmarks [9, 13, 29, 35] evaluate comprehension over pre-segmented clips without any notion of proactive intervention. Streaming comprehension benchmarks [36] assess temporal understanding but do not evaluate proactive behavior. Existing proactive dialogue benchmarks [54, 68] operate under a task-completion setting where the user provides an explicit instruction upfront, and do not model the decision of whether to intervene or remain silent. To fill this gap, we present **EgoServe**, the first large-scale benchmark for proactive assistance in continuous egocentric video. Built upon EgoLife [59], HoloAssist [56], and CaptainCook4D [39], EgoServe spans diverse daily activities, multiple users, and extended temporal contexts ranging from minutes to hours, with proactive services organized into 4 temporal memory horizons and 10 service categories.

On the modeling side, current streaming Video-LLMs lack explicit memory mechanisms for long-horizon retrieval, while retrieval-augmented generation methods [34] have not been applied to the proactive setting where the system must autonomously decide whether a response is warranted. We take a first step with **EgoMemo**, a training-free, memory-augmented agent that maintains three complementary memory representations: (1) multi-scale temporal summaries that hierarchically organize observations from fine-grained clip-level descriptions to coarse activity and session summaries; (2) a semantic knowledge graph encoding entity relationships and activity patterns; and (3) visual embedding archives for similarity-based retrieval. At each timestep, the agent determines whether proactive assistance is warranted and, when deeper context is needed, retrieves and synthesizes information across all three memory stores.

Our contributions are as follows:

- We formalize proactive assistance in continuous video experiences as a decision-driven reasoning task over streaming egocentric perception, and situate it within a taxonomy of three paradigms: reactive, semi-proactive, and proactive.
- We present EgoServe, the first benchmark for evaluating proactive assistance under 4 temporal horizons, comprising over 3,000 service instances across 10 service categories.
- We develop EgoMemo, a training-free, memory-augmented agent that demonstrates the feasibility of proactive assistance through retrieval-augmented reasoning, establishing strong baselines on EgoServe and competitive results on existing egocentric benchmarks.

2 Related Works

Egocentric Video Understanding. Egocentric video understanding has been studied primarily in offline settings where complete recordings are available [11, 12, 16, 17, 19, 21, 22, 30, 41, 42, 45, 47, 55, 61, 67]. Benchmarks such as EgoSchema [35], EgoThink [6] and EgoExoLearn [18], EgoExoBench [15] evaluate episodic memory, cognitive capabilities, and cross-view understanding over pre-recorded footage, while EgoVLP [44] and LaViLa [70] advance video-language pre-training

for egocentric retrieval. A parallel line targets the streaming setting: VideoLLM-online [4], Flash-VStream [65], and StreamChat [32] enable real-time video conversation, Dispider [43] disentangles perception and reasoning for low-latency interaction, and OVO-Bench [36] benchmarks online video comprehension. However, both lines focus on answering questions about observed content, without modeling the decision of whether and when to proactively intervene. Our EgoServe fills this gap by evaluating proactive assistance across multiple temporal memory horizons.

Proactive Large Language Models. Proactive behavior has been explored in language-only settings, where ProAgent [64] enables agents to anticipate teammates' needs in multi-agent cooperation and proactive dialogue systems [7] investigate model-initiated interactions. In the vision-language domain, StreamBridge [54], EWO [69], and ProAssist [68] explore event-triggered proactive generation from streaming video, while Vinci [20] deploys an on-device multimodal proactive assistant. These methods generally conflate event detection with the decision to intervene. In contrast, EgoMemo treats intervention as an explicit reasoning outcome conditioned on retrieved multi-scale historical context.

Proactive Assistive Systems. The vision of context-aware proactive assistance has deep roots in wearable computing. Early work on context-aware applications [46] and contextual awareness in wearable devices [8, 52] established the foundational paradigm of systems that adapt behavior based on sensed user context. Recent advances in foundation models have revived this vision: ContextAgent [58] builds proactive LLM agents on wearable perceptions from smart glasses and earphones, SensibleAgent [26] introduces unobtrusive proactive interaction for AR glasses, and ProAgentBench [53] provides a benchmark for evaluating proactive LLM agents with real-world data. These efforts focus primarily on language-only or short-horizon sensory contexts. In contrast, our work targets proactive assistance over continuous egocentric video streams, requiring long-horizon memory and temporal reasoning across extended activity contexts.

Memory-Based Video Agents. Processing long video within limited context windows has motivated memory-augmented architectures such as MovieChat [51], MA-LMM [14], and VideoAgent [10]. Recent RAG-based approaches [23, 33, 34], VideoRAG [34], Vgent [49], and WorldMM [62], further introduce graph-driven indexing and multi-type memory with adaptive retrieval, but assume offline access to the complete video. EgoMemo departs from these methods in two ways: both memory construction and retrieval are fully streaming, and a VLM-based caption reconstruction step bridges the information gap between structured retrieval results and the contextual descriptions needed for reasoning.

3 EgoServe Benchmark

3.1 Task Formulation

We formulate proactive assistance as a joint decision-and-generation task over continuous video. Let $\mathcal{V} = \{v_1, v_2, \ldots\}$ denote an egocentric video stream segmented into sequential clips. At each timestep t, the agent observes v_t and must

Fig. 2: Overview of the EgoServe benchmark. (a) Annotation pipeline: human annotations from each source dataset are processed through category-specific prompts via a foundation model, followed by manual verification. (b) Response word frequency. (c) Video duration distribution. (d) Per-dataset service counts. (e) Service instance distribution across 10 subcategories and 4 temporal horizons.

produce a binary intervention decision $d_t \in \{0, 1\}$ along with a service response r_t when $d_t = 1$. A correct proactive response requires three conditions to be met: (1) the intervention occurs within a reasonable temporal window of the ground-truth trigger point; (2) the predicted service type matches the ground-truth category; and (3) the generated response is relevant to the identified service need and grounded in the observed context, as assessed by LLM-based evaluation against reference responses.

3.2 Data Source

EgoServe is built upon three egocentric video datasets that together span diverse scenarios and temporal scales. EgoLife [59] provides multi-day continuous daily life recordings, from which we select three participants (A1, A4, A5) across their first five days, enabling evaluation of long-horizon services that require reasoning across temporally distant events, such as connecting observations from different days. HoloAssist [56] captures task-oriented interactions with procedural annotations (step boundaries, error flags, instructor interventions), and we select its 191 validation videos for instant and short-term service evaluation. CaptainCook4D [39] offers structured cooking recordings with both correct and erroneous executions across 24 recipes, from which we select 87 videos with explicit step-error annotations from the validation and test splits, providing a controlled setting for evaluating error detection and corrective guidance.

3.3 Service Taxonomy

A key design principle of EgoServe is that proactive services are organized along two orthogonal dimensions: the temporal memory horizon required to provide the service, and the application context of the service itself. We define 4 temporal memory horizons based on the scope of context the agent must reason over:

- **Instant services** require only the current observation and immediate context, including Safety Alerts (SA: warning about a hazard in the scene) and Tool Use guidance (TU: suggesting a more appropriate tool).
- **Short-Term services** require context spanning the recent minutes of activity, including Error Recovery (ER: detecting and correcting a procedural mistake), Resource Reminder (RR: reminding the user about a recently used resource), and Next-Step Guidance (NSG: suggesting the next action in an ongoing task).
- **Episodic services** require reasoning over the current task, potentially spanning tens of minutes to hours, including Task Reminder (TR: reminding the user of an unfinished task) and Memory Recall (MR: retrieving earlier information that becomes relevant).
- **Long-Term services** require cross-session or multi-day context, including Habit Coaching (HC: suggesting behavioral improvements based on recurring patterns), Routine Optimization (RO: suggesting adjustments to recurring routines), and Memory Link (ML: connecting the current situation to events from previous sessions).

This taxonomy yields 4 major categories and 10 subcategories, as summarized in Table 1. The design reflects a core insight: the difficulty of proactive assistance scales with the temporal horizon of the required context, and a comprehensive benchmark must evaluate across all horizons.

3.4 Annotation Pipeline

Annotating proactive services at scale requires identifying not only what happened in the video, but when assistance would have been appropriate and what the agent should say. We design a semi-automated pipeline that leverages foundation models guided by service category-specific prompts, grounded in existing human annotations from each source dataset.

The annotation pipeline of different datasets is presented in Fig. 2. We leverage the existing human annotations and apply different techniques to different data sources. For HoloAssist, we preprocess the full set of human annotations in chronological order and design tailored prompts that map structured annotations into proactive service instances. For example, instructor corrections map to Error Recovery, and step transitions map to Next-Step Guidance. Due to the task-oriented nature of HoloAssist, annotations primarily cover Instant and Short-Term categories. For EgoLife, we segment annotations into 1-hour intervals and stream them into Gemini together with category-specific prompts. For Instant, Short-Term, and Episodic categories, service dialogues are generated

directly from each interval. For Long-Term services, we adopt a streaming cue-capturing strategy: the model incrementally accumulates events across intervals that may trigger specific long-term service types and continuously generates candidate dialogues. These accumulated cues are then combined with future timeline annotations and re-input into the model, simulating the cross-session reasoning that long-term services require. For CaptainCook4D, we leverage the procedural step annotations and error labels to generate service instances focused on task guidance and error correction. All generated annotations undergo manual verification to ensure temporal accuracy, category correctness, and response quality.

3.5 Evaluation Protocol

EgoServe evaluates proactive assistance along two complementary dimensions:
Temporal precision. For each service category, we match predicted interventions against ground-truth trigger points using a temporal tolerance window δ adapted to the characteristic timescale of each source dataset. A prediction is considered a true positive if its trigger timestamp falls within (start_time$-\delta$, end_time$+\delta$) of a ground-truth service instance *of the same category*. We compute Precision, Recall, and F1 independently for each of the 10 service subcategories.
Response quality. For all successfully matched prediction–ground-truth pairs, we evaluate the quality of the generated response using GPT [38] as an automatic judge, which scores each response on a 1–5 scale across contextual grounding and effectiveness. The LLM-score reports the average over all matched pairs.

4 Methodology

We present EgoMemo, a training-free, memory-augmented agent for proactive assistance in continuous egocentric video. As illustrated in Fig. 3, EgoMemo continuously processes incoming video, maintains structured long-term memory, and performs context-aware reasoning at each timestep. In *proactive* mode, it monitors the evolving scene and autonomously decides when to provide helpful interventions by retrieving relevant historical context. In *reactive* mode, a user query triggers the same retrieval pipeline. Both modes share a unified architecture consisting of two core stages: (1) *streaming memory construction* (Sec. 4.1), which incrementally builds three complementary memory representations, and (2) *streaming retrieval-augmented reasoning* (Sec. 4.2), which retrieves and reconstructs relevant context for decision-making. Both construction and reasoning are conducted incrementally, never requiring access to the complete video.

4.1 Streaming Memory Construction

Given a continuous egocentric video stream $\mathcal{V} = \{v_1, v_2, \ldots\}$, we segment it into non-overlapping short clips. For each clip v_t arriving at timestep t, we generate a dense textual caption c_t using a vision-language model, annotated with its corresponding timestamp, and extract sampled keyframes $\{f_t^1, \ldots, f_t^K\}$. These

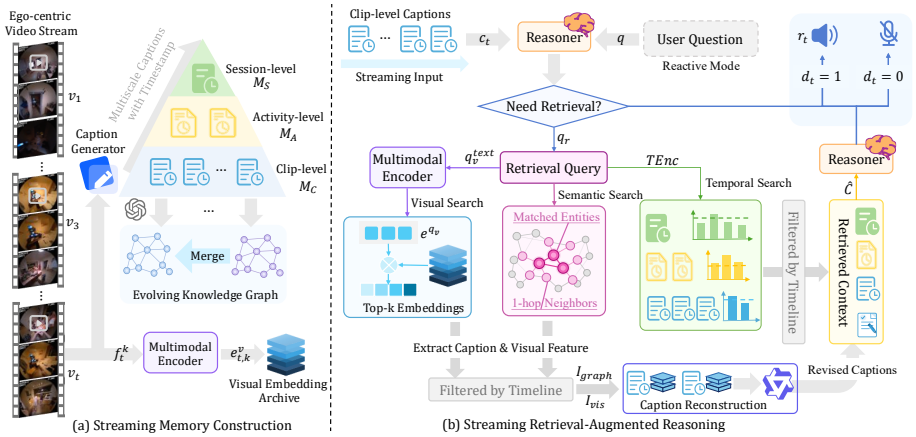

Fig. 3: Architecture of EgoMemo. (a) clip-level captions are incrementally organized into three-level temporal summaries ($\mathcal{M}_C$, $\mathcal{M}_A$, $\mathcal{M}_S$), an evolving knowledge graph $\mathcal{G}$, and a visual embedding archive $\mathcal{A}$. (b) Streaming retrieval-augmented reasoning: three parallel retrieval pathways (temporal, semantic, and visual) gather evidence, which is unified via VLM-based caption reconstruction before the reasoner produces an intervention decision d_t and response r_t.

captions and keyframes are the atomic inputs from which three complementary memory representations are incrementally constructed.

Multi-Scale Temporal Memory. We organize captions into a three-level hierarchy: **Clip-level** captions $\mathcal{M}_C = \{(c_t, t) \mid t = 1, \ldots, T\}$ preserve fine-grained perceptual detail; **Activity-level** summaries $\mathcal{M}_A$ periodically aggregate consecutive clip captions to capture activity context; and **Session-level** summaries $\mathcal{M}_S$ further aggregate activity entries to encode long-horizon routines. Formally:

$$M_A^{(j)} = \textbf{Summarize}\big(\{c_t\}_{t \in \mathcal{W}_j}\big), \quad M_S^{(k)} = \textbf{Summarize}\Big(\{M_A^{(j)}\}_{j \in \mathcal{W}_k}\Big), \quad (1)$$

where $\mathcal{W}_j$ and $\mathcal{W}_k$ denote temporal windows at the Activity and Session levels, respectively. The roll-up is fully incremental: only newly accumulated segments trigger summarization at the next level. To support retrieval, we encode captions at all three levels into dense embeddings $\{e_i^c, e_j^a, e_k^s\}$ using a text encoder `TEnc` and index them for similarity search.

Evolving Knowledge Graph. To capture semantic relationships between entities across different time segments, we maintain a knowledge graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ that evolves as new observations arrive. For each caption c_t, we prompt an LLM to extract entities and relations $\mathcal{R}_t = \textbf{Extract}(c_t)$, which are merged into the global graph through name-based entity resolution: $\mathcal{G}_t = \textbf{Merge}(\mathcal{G}_{t-1}, \mathcal{R}_t)$. Each node in $\mathcal{G}$ maintains links to its source captions, enabling the graph to serve as a structured index over the temporal memory.

Visual Embedding Archive. To complement text-based memories with visual details that are difficult to verbalize (object appearances, spatial layouts), we

encode sampled keyframes using a multimodal encoder and store the resulting embeddings alongside their source caption index and timestamp:

$$\mathbf{e}^{v}_{t,k} = \mathbf{MEnc}(f^{k}_{t}) \in \mathbb{R}^{d_v}, \quad \mathcal{A} = \{(\mathbf{e}^{v}_{t,k},\ t) \mid t = 1, \ldots, T;\ k = 1, \ldots, K\}. \quad (2)$$

This enables similarity-based retrieval of visually relevant moments that may lack lexical overlap with the query.

4.2 Streaming Retrieval-Augmented Reasoning

At each timestep t, the LLM reasoning agent receives the current clip-level caption c_t and the most recent short-term context. For proactive assistance, it assesses whether the current observation warrants an active intervention; for reactive QA, a user query q serves as an external trigger with $d_t = 1$ by default. In both cases, when deeper context is needed, the agent generates a retrieval query q_r and invokes three parallel retrieval pathways, followed by a caption reconstruction step that synthesizes the retrieved evidence into a coherent context for final reasoning.

Multi-Scale Temporal Retrieval. We search the multi-scale temporal memory by encoding q_r into a dense embedding and retrieving the top-k similar captions:

$$\mathcal{I}_{\text{temp}} = \underset{t'}{\text{top-k}}\, \text{sim}(\mathbf{TEnc}(q_r),\ \mathbf{e}^{c}_{t'}), \quad t' \leq t. \quad (3)$$

The constraint $t' \leq t$ enforces the streaming setting. Clip-level entries provide fine-grained evidence, while activity- and session-level summaries offer broader context for queries involving extended activities or recurring patterns.

Graph-Based Semantic Retrieval. We extract key entities from q_r, match them against nodes in $\mathcal{G}$, and expand to their 1-hop neighbors to capture semantically associated events:

$$\mathcal{N}_{\text{match}} = \mathbf{EntityMatch}(q_r,\ \mathcal{G}), \qquad \mathcal{I}_{\text{graph}} = \bigcup_{n \in \mathcal{N}_{\text{match}}} \mathbf{Neighbor}(n,\ \mathcal{G}), \quad (4)$$

where $\mathcal{I}_{\text{graph}}$ collects caption indices linked to the matched nodes and their neighbors. This captures events described with different wording or occurring at distant timesteps.

Visual Similarity Retrieval. Since q_r is in text form, we first rewrite it into a visual-centric description q^{text}_{v} emphasizing visual attributes, encode it via the multimodal encoder to obtain $\mathbf{e}^{q_v} = \mathbf{MEnc}(q^{\text{text}}_{v})$, and retrieve the top-$k$ matching keyframe embeddings:

$$\mathcal{I}_{\text{vis}} = \underset{t'}{\text{top-k}}\, \text{sim}(\mathbf{e}^{q_v},\ \mathbf{e}^{v}_{t'}), \quad t' \leq t. \quad (5)$$

Each retrieved visual embedding is linked to its source clip-level caption, yielding a set of caption indices $\mathcal{I}_{\text{vis}}$.

Caption Reconstruction and Reasoning. Both graph-based and visual retrieval return sets of caption indices rather than self-contained descriptions. To

bridge this gap, we apply a unified *VLM-based caption reconstruction* step: for each set of retrieved indices $\mathcal{I} \in \{\mathcal{I}_{\text{graph}}, \mathcal{I}_{\text{vis}}\}$, we gather the corresponding clip-level captions and keyframes, and prompt a VLM to generate a reconstructed caption conditioned on the retrieval query:

$$\hat{c}_{\mathcal{I}} = \textbf{VLM\_Reconstruct}(\{(c_{t'}, f_{t'})\}_{t' \in \mathcal{I}},\ q_r)\,. \tag{6}$$

This reconstruction resolves co-references, recovers visual details absent from the original captions, and produces a query-focused narrative. The outputs from the temporally retrieved captions $\{c_{t'}\}_{t' \in \mathcal{I}_{\text{temp}}}$, graph-reconstructed context $\hat{c}_{\text{graph}}$, and visually reconstructed context $\hat{c}_{\text{vis}}$, are aggregated into a unified retrieved context $\hat{\mathcal{C}}$ and provided to the LLM reasoning agent:

$$(d_t, r_t) = \textbf{Reason}(c_t,\ \hat{\mathcal{C}},\ q_r), \tag{7}$$

where $d_t \in \{0, 1\}$ is the intervention decision and r_t is the generated response. For reactive QA, the same pipeline is invoked with the user query as q_r and $d_t = 1$. This unified formulation requires no architectural modification between the two modes.

4.3 Adaptation to Other Benchmarks

While EgoMemo is designed for streaming scenarios, its architecture naturally accommodates offline video understanding tasks where the full video is available at inference time. In this setting, we first process the entire video sequentially to construct the complete memory representations ($\mathcal{M}_C$, $\mathcal{M}_A$, $\mathcal{M}_S$, $\mathcal{G}$, and $\mathcal{A}$) in a single forward pass. Given a query q, the agent adopts a coarse-to-fine perception strategy: it first receives all activity-level and session-level summaries ($\mathcal{M}_A$ and $\mathcal{M}_S$) to form a global understanding of the video, and attempts to answer the question based on this high-level context alone. If the agent determines that the available information is insufficient, it invokes the same retrieval pipeline described in Sec. 4.2 to retrieve fine-grained clip-level evidence from $\mathcal{M}_C$, $\mathcal{G}$, and $\mathcal{A}$. This coarse-to-fine strategy avoids unnecessary retrieval for questions answerable from high-level summaries, while preserving access to detailed evidence when needed. The adaptation requires no modification to the model architecture, demonstrating that the streaming-first design of EgoMemo generalizes gracefully to offline settings. We evaluate both streaming and offline performance in Sec. 5.

5 Experiments

5.1 Experimental Setup

Evaluation Datasets We evaluate EgoMemo across three benchmark categories. (1) **Proactive assistance**: our EgoServe benchmark, which evaluates proactive service triggering over streaming egocentric video. (2) **Online video understanding**: ESTP-Bench [69], which assesses ego-proactive video understanding through temporally grounded questions at opportune moments; and

Table 1: Evaluation results on EgoServe benchmark.

Model	Instant		Short-term			Episodic		Long-term			Overall	LLM-score
	SA	*TU*	*NSG*	*ER*	*RR*	*MR*	*TR*	*HC*	*ML*	*RO*		
Qwen3-VL-Plus [2]	5.2	1.5	8.6	**10.4**	3.4	0.0	1.8	4.4	0.0	0.0	3.5	2.8
GPT-5-mini [38]	**12.5**	3.6	9.5	1.0	5.2	0.0	**9.4**	**5.7**	0.0	0.0	4.7	**3.0**
w/o MS	11.2	8.0	24.2	1.9	4.9	**4.8**	2.9	4.6	1.9	5.7	7.0	2.8
w/o Recons.	10.5	7.6	**26.1**	0.4	**5.8**	0.0	4.8	5.4	0.0	5.7	6.6	2.8
w/o VSR	11.7	6.0	24.8	1.2	3.9	1.2	6.6	4.2	3.2	6.4	6.9	2.8
w/o GSR	**12.5**	5.6	24.4	1.9	4.9	2.6	4.8	2.4	0.0	5.4	6.5	2.8
w/o MTR	9.7	**8.8**	24.6	1.2	4.0	4.0	4.5	1.9	2.1	7.5	6.8	2.7
EgoMemo (Ours)	11.4	7.5	24.7	1.7	4.7	3.8	5.7	3.7	**4.9**	**11.8**	**8.0**	2.8

OVO-Bench [36], which evaluates online video comprehension across backward tracing, real-time perception over 858 videos. (3) **Offline egocentric QA**: EgoSchema [35], a long-form temporal reasoning benchmark with over 5,000 multiple-choice questions spanning 250+ hours of Ego4D video; EgoTaskQA [24], which tests understanding of task-oriented egocentric activities, converted to multiple-choice format following [10]; and QAEgo4D [3], which evaluates episodic memory querying over Ego4D recordings.

Implementation Details For streaming memory construction, we use Qwen3-VL-8B-Instruct as the VLM to generate clip-level captions $\mathcal{M}_L$ from each video segment. Activity-level summaries $\mathcal{M}_A$ and Session-level summaries $\mathcal{M}_S$ are produced via LLM-based summarization. The temporal windows $\mathcal{W}_j$ and $\mathcal{W}_k$ are adapted to the video duration: for ultra-long recordings such as EgoLife, we use 30s/5min/1h for clip/activity/session levels, while for shorter online benchmarks we use 10s/1min/5min. Entity and relation extraction for the evolving knowledge graph is performed using GPT-4o-mini. For retrieval, we encode textual captions with OpenAI's text-embedding-3-small as the text encoder `TEnc`, and encode keyframes with ImageBind as the multimodal encoder `MEnc`. The reasoning agent uses GPT-5-mini for online streaming benchmarks and GPT-5.2 for offline benchmarks. For EgoServe evaluation, the temporal tolerance window δ is set per dataset according to its characteristic timescale: $\delta = 60$s for EgoLife, $\delta = 25$s for CaptainCook4D, and $\delta = 10$s for HoloAssist. Further details on hyperparameters and prompt designs are provided in the supplementary material.

5.2 Results on EgoServe

Main Comparison Table 1 reports per-category F1 scores on the EgoServe benchmark, where a prediction counts as a true positive only if it falls within the dataset-specific temporal window of a ground-truth instance of the *same* service category. We compare EgoMemo against two strong proprietary models: GPT-5-mini and Qwen3-VL-Plus. Both models struggle substantially: Qwen fails almost entirely on episodic and long-term services where historical context is essential. GPT-5-mini performs better (4.7 overall), particularly on instant

Table 2: Experimental results of various models evaluated on the ESTP-Bench.

Model	Explicit Proactive Task									Implicit Proactive Task				
	OR	*AP*	*TRU*	*OL*	*OSC*	*EOL*	*EOSC*	*AR*	*All*	*OFR*	*IFR*	*NAR*	*TU*	*All*
EyeWO*	26.6	26.6	25.1	26.8	19.8	22.3	20.8	20.7	23.6	24.8	31.0	75.3	78.7	52.5
LLaVA-OneVision [27]	8.3	8.8	22.8	25.4	13.5	9.8	9.6	10.3	13.6	20.3	20.9	35.9	49.9	31.8
Qwen2-VL [1]	13.7	13.5	15.4	**29.5**	8.0	15.4	16.6	10.9	15.4	17.8	19.8	**56.4**	63.1	**39.3**
MiniCPM-V [60]	14.9	16.8	17.1	26.8	7.7	12.9	12.5	13.1	15.2	15.9	21.0	46.8	62.2	36.5
LLaVA-NeXT-Video [27]	15.6	14.6	21.9	26.8	12.8	14.2	13.5	12.3	16.5	18.6	23.2	44.9	51.6	34.6
InternVL-V2 [5]	11.3	5.9	7.0	10.1	0.7	2.7	5.2	2.2	5.6	8.3	2.9	4.3	11.2	6.7
LIVE [4]	11.2	13.9	7.9	13.2	5.6	9.4	6.0	8.9	9.5	5.8	8.9	41.0	46.7	25.6
MMDuet [57]	7.2	10.3	17.6	10.2	4.2	6.1	8.8	8.5	9.1	10.0	7.7	50.1	**69.1**	34.2
EgoMemo (Ours)	**25.4**	**30.5**	**32.4**	22.9	**26.6**	**26.1**	**35.8**	**21.1**	**27.6**	**23.8**	**29.9**	36.7	48.4	34.7

Safety alerts (12.5) and short-term Next-Step Guidance (9.5), but still scores zero on all long-term categories except Habit Coaching (5.7), confirming that even powerful VLMs cannot provide meaningful proactive assistance without access to accumulated context.

EgoMemo achieves 8.0 overall F1, nearly doubling GPT-5-mini's performance. The gains are most pronounced on long-term services: Memory Link reaches 4.9 (vs. 0.0 for both baselines) and Routine Optimization reaches 11.8 (vs. 0.0), demonstrating that our structured memory and retrieval mechanisms successfully enable cross-session reasoning. On episodic services, EgoMemo attains non-zero scores on both Memory Recall (3.8) and Task Reminder (5.7), whereas the baselines fail on at least one. Short-term services show strong performance across Next-Step Guidance (24.7) and Resource Reminder (4.7). For matched predictions, the LLM-score reaches 2.8. We note that GPT-5-mini matches far fewer service instances overall (4.7 vs. 8.0), indicating that the primary bottleneck for baselines lies in temporal precision rather than response quality.

The absolute scores remain moderate across all methods, reflecting the genuine difficulty of proactive assistance: the agent must simultaneously decide *when* to intervene within the correct temporal window, identify the appropriate service type, and produce a helpful response. EgoServe is designed to expose these challenges and serve as a diagnostic tool for future progress.

Ablation Study We ablate each component of EgoMemo on EgoServe to understand its individual contribution. Results are reported in Table 1.

Replacing the three-level temporal hierarchy with a single-scale caption store (*w/o MS*) reduces the overall F1 from 8.0 to 7.0, with Memory Link dropping from 4.9 to 1.9 and Routine Optimization from 11.8 to 5.7, confirming that the hierarchical organization of clip-, activity-, and session-level summaries is essential for capturing patterns across extended temporal spans. Removing the VLM-based caption reconstruction step (*w/o Recons.*) causes a large overall drop (8.0 → 6.6). Both Memory Recall and Memory Link fall to 0.0, indicating that raw retrieved snippets are insufficient for the reasoner to synthesize coherent cross-temporal evidence. Disabling the visual embedding archive (*w/o VSR*)

Table 3: Evaluation results on OVO-Bench.

Model	Real-Time Visual Perception							Backward Tracing			
	OCR	*ACR*	*ATR*	*STU*	*FPD*	*OJR*	*Avg.*	*EPM*	*ASI*	*HLD*	*Avg.*
GPT-4o [37]	69.8	64.22	71.55	51.12	70.3	59.78	64.46	**57.91**	**75.68**	48.66	**60.75**
Qwen2-VL-72B [1]	65.77	60.55	69.83	51.69	69.31	54.35	61.92	52.53	60.81	**57.53**	56.95
LLaVA-Video-7B [27]	69.13	58.72	68.83	49.44	**74.26**	59.78	63.52	56.23	57.43	7.53	40.4
LLaVA-OneVision-7B [27]	66.44	57.80	73.28	53.37	71.29	61.96	64.02	54.21	55.41	21.51	43.71
InternVL-V2-8B [5]	67.11	60.55	63.79	46.07	68.32	56.52	60.39	48.15	57.43	24.73	43.44
LongVU-7B [48]	53.69	53.21	62.93	47.75	68.32	59.78	57.61	40.74	59.46	4.84	35.01
VideoLLM-online-8B [4]	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73
EyeWO [69]	24.16	27.52	31.89	32.58	44.55	35.87	32.76	39.06	38.51	6.45	28.00
Dispider [43]	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	4.3	36.06
EgoMemo (Ours)	**91.28**	**75.23**	**77.59**	**62.92**	69.31	**75.54**	**75.15**	52.53	50.00	44.62	49.60

yields a modest reduction (8.0 → 6.9), with the impact concentrated on Memory Link (4.9 → 3.2) and Routine Optimization (11.8 → 6.4), where visual cues help identify recurring objects or scenes that textual descriptions alone may miss. Removing the knowledge graph pathway (*w/o GSR*) reduces F1 to 6.5, with Memory Link dropping sharply to 0.0, highlighting the graph's role in linking semantically related entities across time. Without multi-scale temporal retrieval (*w/o MTR*), F1 drops to 6.8, and Memory Link falls to 2.1, as the system loses its primary mechanism for locating temporally relevant context at the appropriate granularity.

Taken together, caption reconstruction and graph semantic retrieval have the largest individual impact, while all three retrieval pathways (temporal, semantic, and visual) contribute complementary evidence. No single pathway subsumes another, validating the design of parallel heterogeneous retrieval for proactive assistance.

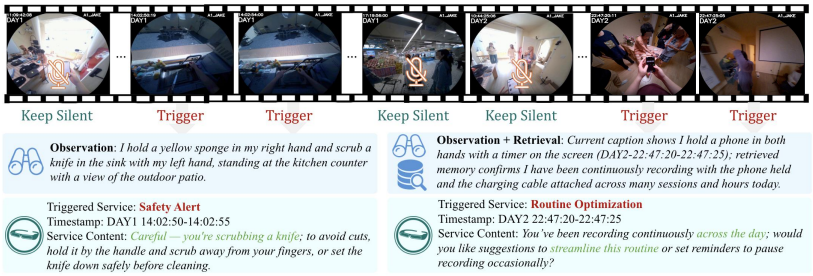

Fig. 4: Qualitative examples of EgoMemo's proactive assistance. Left: an instant Safety Alert service triggered by observing the user scrubbing a knife barehanded. Right: a long-term Routine Optimization service triggered on Day 2 by detecting a recurring pattern of prolonged phone recording across multiple sessions and days.

Qualitative Analysis Fig. 4 shows two representative examples of EgoMemo's proactive behavior on EgoServe. In the left example, the agent observes the user scrubbing a knife and immediately triggers a Safety Alert based solely on the current clip-level caption. In the right example, the agent detects on Day 2 that the user has been holding a phone with a timer on screen; it retrieves cross-session memory confirming a recurring pattern of prolonged recording across multiple days, and triggers a Routine Optimization service suggesting ways to streamline this behavior.

5.3 Generalization to Existing Benchmarks

We further evaluate EgoMemo on five established benchmarks to verify that its architecture generalizes beyond proactive assistance.

ESTP-Bench (Table 2). EgoMemo achieves the best overall score of 27.6 on explicit proactive tasks, surpassing EyeWO (23.6), with particularly strong gains on context-dependent subtasks such as TRU (32.4 vs. 25.1) and EOSC (35.8 vs. 20.8). On implicit proactive tasks, EgoMemo scores 34.7 compared to EyeWO's 52.5. Notably, EyeWO is the only trained model in this comparison; all others, including EgoMemo, are entirely training-free.

Table 4: Results on offline egocentric video benchmarks. For EgoTaskQA, we convert the dataset into MCQ format following [66].

Method	*Ego-TaskQA*	*QA-Ego4D*	*Ego-Schema*
LLaVA-OneVision-7B [27]	55.8	65.7	60.1
VideoLLaMA3 [63]	56.6	62.4	61.1
VideoChat2-HD [29]	45.5	52.0	55.8
Exo2Ego-7B [66]	48.1	62.1	61.3
Qwen2-VL-7B [1]	57.9	60.3	63.3
EgoThinker [40]	**64.4**	66.2	67.6
EgoMemo (Ours)	60.9	**68.0**	**74.8**

OVO-Bench (Table 3). EgoMemo achieves the best real-time perception score of 75.15, outperforming GPT-4o (64.46) and LLaVA-OneVision (64.02). On backward tracing, EgoMemo scores 49.60, competitive with GPT-4o (60.75); the gap is expected given the substantially larger capacity and context windows of proprietary models.

Offline egocentric QA (Table 4). EgoMemo achieves the best scores on two of the three egocentric QA datasets: EgoSchema (74.8, +7.2 over EgoThinker) and QAEgo4D (68.0 vs. 66.2), demonstrating that structured memory access is particularly effective for long-form temporal reasoning. On EgoTaskQA, EgoMemo scores 60.9, competitive with EgoThinker (64.4), where fine-grained action-state reasoning within short procedural segments favors direct visual perception over memory retrieval.

These results confirm that EgoMemo's streaming-first design generalizes across both streaming and offline settings without architectural modification.

6 Conclusion

We presented EgoServe, the first large-scale benchmark for proactive assistance in continuous egocentric video, comprising over 3,000 service instances across 10 categories and 4 temporal memory horizons. Alongside the benchmark, we

introduced EgoMemo, a training-free, memory-augmented agent that maintains multi-scale temporal summaries, an evolving knowledge graph, and a visual embedding archive to perform retrieval-augmented reasoning over streaming video. Experiments show that EgoMemo establishes strong baselines on EgoServe and achieves competitive or state-of-the-art results on 5 existing benchmarks.

Limitations and future work. EgoMemo's text-based memory loses fine-grained visual details during captioning, and the name-based entity resolution may fail for visually ambiguous entities. The semi-automated annotation pipeline may also introduce biases toward service types that are easier for foundation models to generate. Future directions include learning intervention timing from human preferences, incorporating audio context, and extending EgoServe to multi-turn proactive dialogues and multi-user settings.

7 Acknowledgement

The paper is supported in part by the National Natural Science Foundation of China under grant No.62441231, 62293542, Liao Ning Province Science and Technology Plan No.2023JH26/10200016, Dalian City Science and Technology Innovation Fund No.2023JJ11CG001, and Ningbo Key RD project under Grant No.2025Z039.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 12, 13, 14
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 11
3. Bärmann, L., Waibel, A.: Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1560–1568 (2022) 11
4. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18407–18418 (2024) 4, 12, 13
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 12, 13
6. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024) 3
7. Deng, Y., Lei, W., Lam, W., Chua, T.S.: A survey on proactive dialogue systems: Problems, methods, and prospects. arXiv preprint arXiv:2305.02750 (2023) 4

8. Dey, A.K.: Understanding and using context. Personal and ubiquitous computing **5**(1), 4–7 (2001) 4
9. Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J.: Benchmarking robustness of 3d object detection to common corruptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1022–1032 (2023) 3
10. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024) 4, 11
11. Girdhar, R., Grauman, K.: Anticipative video transformer. In: ICCV (2021) 3
12. Goyal, M., Modi, S., Goyal, R., Gupta, S.: Human hands as probes for interactive object understanding. In: CVPR (2022) 3
13. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022) 3
14. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13504–13514 (2024) 4
15. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first-and third-person view video understanding in mllms. arXiv preprint arXiv:2507.18342 (2025) 3
16. Huang, Y., Cai, M., Li, Z., Lu, F., Sato, Y.: Mutual context network for jointly estimating egocentric gaze and action. IEEE Transactions on Image Processing **29**, 7795–7806 (2020) 3
17. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: ECCV (2018) 3
18. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al.: Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22072–22086 (2024) 3
19. Huang, Y., Sugano, Y., Sato, Y.: Improving action segmentation via graph-based temporal reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14024–14034 (2020) 3
20. Huang, Y., Xu, J., Pei, B., He, Y., Chen, G., Yang, L., Chen, X., Wang, Y., Nie, Z., Liu, J., et al.: Vinci: A real-time embodied smart assistant based on egocentric vision-language model. arXiv preprint arXiv:2412.21080 (2024) 2, 4
21. Huang, Y., Yang, L., Chen, G., Zhang, H., Lu, F., Sato, Y.: Matching compound prototypes for few-shot action recognition. International Journal of Computer Vision **132**(9), 3977–4002 (2024) 3
22. Huang, Y., Yang, L., Sato, Y.: Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18908–18918 (2023) 3
23. Jeong, S., Kim, K., Baek, J., Hwang, S.J.: Videorag: Retrieval-augmented generation over video corpus. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 21278–21298 (2025) 4
24. Jia, B., Lei, T., Zhu, S.C., Huang, S.: Egotaskqa: Understanding human tasks in egocentric videos. Advances in Neural Information Processing Systems **35**, 3343–3360 (2022) 11

25. Kang, C., Huang, Y., Ouyang, L., Zhang, M., Liu, R., Sato, Y.: Can mllms read the room? a multimodal benchmark for assessing deception in multi-party social interactions. arXiv preprint arXiv:2511.16221 (2025) 2
26. Lee, G., Xia, M., Numan, N., Qian, X., Li, D., Chen, Y., Kulshrestha, A., Chatterjee, I., Zhang, Y., Manocha, D., et al.: Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–22 (2025) 4
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 12, 13, 14
28. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025) 2
29. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024) 3, 14
30. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexofitness: Towards egocentric and exocentric full-body action understanding. In: European conference on computer vision. pp. 363–382. Springer (2024) 3
31. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024) 2
32. Liu, J., Yu, Z., Lan, S., Wang, S., Fang, R., Kautz, J., Li, H., Alvare, J.M.: Streamchat: Chatting with streaming video. arXiv preprint arXiv:2412.08646 (2024) 4
33. Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., Li, W.: Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. arXiv preprint arXiv:2508.09736 (2025) 4
34. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., et al.: Video-rag: Visually-aligned retrieval-augmented long video comprehension. arXiv preprint arXiv:2411.13093 (2024) 3, 4
35. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023) 3, 11
36. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18902–18913 (2025) 3, 4, 11
37. OpenAI: Introducing GPT-4o and More Tools to ChatGPT Free Users. https://openai.com (2024), accessed: 2026-03-06 13
38. OpenAI: GPT-5.1 (2025), https://openai.com, large language model 7, 11
39. Peddi, R., Arya, S., Challa, B., Pallapothula, L., Vyas, A., Gouripeddi, B., Zhang, Q., Wang, J., Komaragiri, V., Ragan, E., et al.: Captaincook4d: A dataset for understanding errors in procedural activities. Advances in Neural Information Processing Systems **37**, 135626–135679 (2024) 3, 5
40. Pei, B., Huang, Y., Xu, J., He, Y., Chen, G., Wu, F., Qiao, Y., Pang, J.: Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. arXiv preprint arXiv:2510.23569 (2025) 14

41. Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G.M., Damen, D., Tommasi, T.: An outlook into the future of egocentric vision. arXiv preprint arXiv:2308.07123 (2023) 3
42. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: E2 (go) motion: Motion augmented event stream for egocentric action recognition. In: CVPR (2022) 3
43. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24045–24055 (2025) 2, 4, 13
44. Qinghong Lin, K., Jinpeng Wang, A., Soldan, M., Wray, M., Yan, R., Zhongcong Xu, E., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. arXiv e-prints pp. arXiv–2206 (2022) 2, 3
45. Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., Tuytelaars, T.: Multimodal distillation for egocentric action recognition. In: ICCV (2023) 3
46. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: 1994 first workshop on mobile computing systems and applications. pp. 85–90. IEEE (1994) 4
47. Shan, D., Geng, J., Shu, M., Fouhey, D.: Understanding human hands in contact at internet scale. In: CVPR (2020) 3
48. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024) 2, 13
49. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. arXiv preprint arXiv:2510.14032 (2025) 4
50. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26160–26169 (2025) 2
51. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024) 2, 4
52. Starner, T.: Wearable computing and contextual awareness. Ph.D. thesis, Massachusetts Institute of Technology (1999) 4
53. Tang, Y., Tang, H., Cao, T., Nguyen, L., Zhang, A., Cao, X., Liu, C., Ding, W., Li, Y.: Proagentbench: Evaluating llm agents for proactive assistance with real-world data. arXiv e-prints pp. arXiv–2602 (2026) 4
54. Wang, H., Feng, B., Lai, Z., Xu, M., Li, S., Ge, W., Dehghan, A., Cao, M., Huang, P.: Streambridge: Turning your offline video large language model into a proactive streaming assistant. arXiv preprint arXiv:2505.05467 (2025) 2, 3, 4
55. Wang, H., Singh, M.K., Torresani, L.: Ego-only: Egocentric action detection without exocentric transferring. In: ICCV (2023) 3
56. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023) 3, 5

57. Wang, Y., Meng, X., Wang, Y., Liang, J., Wei, J., Zhang, H., Zhao, D.: Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. arXiv preprint arXiv:2411.17991 **1**(3), 5 (2024) 12
58. Yang, B., Xu, L., Zeng, L., Liu, K., Jiang, S., Lu, W., Chen, H., Jiang, X., Xing, G., Yan, Z.: Contextagent: Context-aware proactive llm agents with open-world sensory perceptions. arXiv preprint arXiv:2505.14668 (2025) 4
59. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025) 2, 3, 5
60. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 12
61. Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., Zhang, B., You, H., Xu, D., Gan, Z., et al.: MM-Ego: Towards building egocentric multimodal llms for video qa. In: International Conference on Learning Representations (ICLR) (2025) 3
62. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. arXiv preprint arXiv:2512.02425 (2025) 4
63. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025) 14
64. Zhang, C., Yang, K., Hu, S., Wang, Z., Li, G., Sun, Y., Zhang, C., Zhang, Z., Liu, A., Zhu, S.C., et al.: Proagent: building proactive cooperative agents with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17591–17599 (2024) 2, 4
65. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Dai, J., Jin, X.: Flash-vstream: Memory-based real-time understanding for long video streams. arXiv preprint arXiv:2406.08085 (2024) 2, 4
66. Zhang, H., Chu, Q., Liu, M., Shi, H., Wang, Y., Nie, L.: Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. arXiv preprint arXiv:2503.09143 (2025) 14
67. Zhang, L., Zhou, S., Stent, S., Shi, J.: Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In: ECCV (2022) 3
68. Zhang, Y., Dong, X.L., Lin, Z., Madotto, A., Kumar, A., Damavandi, B., Chai, J., Moon, S.: Proactive assistant dialogue generation from streaming egocentric videos. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12055–12079 (2025) 2, 3, 4
69. Zhang, Y., Shi, C., Wang, Y., Yang, S.: Eyes wide open: Ego proactive video-llm for streaming video. arXiv preprint arXiv:2510.14560 (2025) 2, 4, 10, 13
70. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6586–6597 (2023) 3
71. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., et al.: Apollo: An exploration of video understanding in large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18891–18901 (2025) 2