EgoLife: Towards Egocentric Life Assistant

Jingkang Yang[1,2], Shuai Liu[1,2], Hongming Guo[3], Yuhao Dong[1], Xiamengwei Zhang[3], Sicheng Zhang[4], Pengyun Wang[3], Zitang Zhou[3], Binzhu Xie[3], Ziyue Wang[1], Bei Ouyang[5], Zhengyu Lin[1], Marco Cominelli[7], Zhongang Cai[1], Bo Li[1,2], Yuanhan Zhang[1,2], Peiyuan Zhang[1,2], Fangzhou Hong[1], Joerg Widmer[5], Francesco Gringoli[6], Lei Yang[1], Ziwei Liu[1,2,✉]

[1] S-Lab, Nanyang Technological University, Singapore [2] LMMs-Lab
[3] Beijing University of Posts and Telecommunications, China [4] Capital Normal University, China
[5] IMDEA Networks, Spain [6] University of Brescia, Italy [7] Politecnico di Milano, Italy

https://egolife-ai.github.io/

Figure 1. **The Overview of EgoLife Project.** The EgoLife project features six participants living together for a week to prepare an Earth Day celebration. Each participant wears Meta Aria glasses [1], recording approximately 8 hours of egocentric video and signals daily. In addition, 15 cameras and 2 mmWave devices provide synchronized third-person perspective data (detailed in Figure 2). These comprehensive annotations enable the development of state-of-the-art multimodal egocentric AI assistants and introduce novel tasks to advance long-term egocentric life assistance, as illustrated in the EgoLife task board.

Abstract

*We introduce **EgoLife**, a project to develop an egocentric life assistant that accompanies and enhances personal efficiency through AI-powered wearable glasses. To lay the foundation for this assistant, we conducted a comprehensive data collection study where six participants lived together for one week, continuously recording their daily activities—including discussions, shopping, cooking, socializing, and entertainment—using AI glasses for multimodal egocentric video capture, along with synchronized third-person-view video references. This effort resulted in the **EgoLife Dataset**, a comprehensive 300-hour egocentric, interpersonal, multiview, and multimodal daily life dataset with intensive annotation. Leveraging this dataset, we introduce EgoLifeQA, a suite of long-context, life-oriented question-answering tasks designed to provide meaningful assistance in daily life by addressing practical questions such as recalling past relevant events, monitoring health habits, and offering personalized recommendations.*

*To address the key technical challenges of **1)** developing robust visual-audio models for egocentric data, **2)** enabling*

✉Corresponding author: Ziwei Liu. Full author list is in Appendix A.

*identity recognition, and **3)** facilitating long-context question answering over extensive temporal information, we introduce **EgoBulter**, an integrated system comprising **EgoGPT** and **EgoRAG**. EgoGPT is an omni-modal model trained on egocentric datasets, achieving state-of-the-art performance on egocentric video understanding. EgoRAG is a retrieval-based component that supports answering ultra-long-context questions. Our experimental studies verify their working mechanisms and reveal critical factors and bottlenecks, guiding future improvements. By releasing our datasets, models, and benchmarks, we aim to stimulate further research in egocentric AI assistants.*

1. Introduction

Imagine a future where an AI assistant seamlessly integrates into daily life, offering personalized food suggestions based on your habits and reminding you of purchases made after work, all through a comprehensive analysis of your potential needs not only from your activities but also those of your family. Such an assistant would greatly enhance both personal and interpersonal efficiency, offering meaningful, life-oriented assistance and delivering actionable insights. Realizing this vision requires significant advancements in understanding ultra-long-term behavior patterns and the intricate dynamics of social interactions—areas where current egocentric vision systems and datasets still fall short [2, 3].

While existing datasets like Epic-Kitchen [4] and Ego4D [5] support numerous valuable tasks, they are limited by relatively short recording durations and a predominantly monographic perspective. These limitations hinder their ability to capture comprehensive habits and the intricate dynamics of social interactions. Overcoming these challenges requires a dataset that spans extended activities, integrates multimodal data, and incorporates multi-person perspectives to reflect the complexity of real-life experiences.

In response to these challenges, we initiated *the Project EgoLife*. As shown in Figure 1, over one week, six participants shared a fully instrumented living environment, recording approximately eight hours of egocentric multimodal video daily using Meta Aria glasses [1]. This resulted in the **EgoLife dataset**, a rich 300-hour collection of egocentric, multimodal, and multi-view data, augmented with synchronized third-person perspectives captured from 15 additional cameras [6] and two mmWave devices [7] (see Figure 2 showing their arrangements). The dataset provides an unprecedented resource for studying long-duration activities, interpersonal dynamics, and contextual interactions, with rich annotations including audio transcript and visual-audio narrations at various time granularity.

Building on the EgoLife dataset, we introduce the **EgoLifeQA benchmark**, a set of long-context, life-oriented question-answering tasks that assess the effectiveness of personalized AI assistance. These tasks address practical, everyday needs such as locating misplaced items, recalling past events, tracking health habits, analyzing social interactions, and making timely recommendations. By enabling context-aware responses to questions like “Where are the scissors, and who used them last?”, “How much water did I consume today?”, or “Based on today’s consumption, what should I purchase or restock later?”, EgoLifeQA aims to inspire methods that provide intelligent, anticipatory support, simplifying daily activities and enhancing the user experience.

Addressing the novel tasks posed by the EgoLifeQA requires innovative technical contributions to tackle key challenges: **1)** developing robust omni-modal models that integrate both visual and audio data specifically for egocentric contexts, **2)** achieving accurate recognition and tracking of individuals, and **3)** enabling ultra-long-context (week-level) question answering over extensive temporal sequences. To meet these objectives, we present **EgoButler**, an integrated system comprising *EgoGPT*, a lightweight personalized vision-audio-language model fine-tuned on egocentric datasets for state-of-the-art multimodal video understanding, and *EgoRAG* - a retrieval-augmented generation module supports long-context question answering. Our comprehensive evaluations identify crucial factors and highlight existing bottlenecks, offering valuable insights and paving the way for future advancements in egocentric life AI assistance.

In sum, the EgoLife project contributes a comprehensive *EgoLife dataset*, *EgoLifeQA tasks*, and the *EgoButler system*, addressing key challenges in egocentric AI by enabling long-context understanding, multimodal integration, and personalized assistance. These resources fill critical gaps left by existing datasets and models, laying a robust foundation for future research on life-oriented AI. Looking ahead, we plan to expand the dataset to cover a broader range of languages, locations, and activities, and develop more sophisticated models that push the boundaries of AI’s ability to understand and enhance everyday life. Ultimately, we aim to move closer to a world where AI glasses seamlessly support and enrich the human experience.

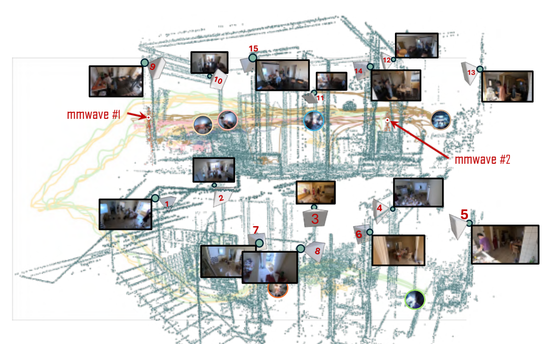


Figure 2. 3D reconstruction of the shared house using Aria Multi-MPS [1], showcasing the locations of 15 Exo cameras in the common area and 2 mmWave devices (highlighted in red) on the second floor. Color-coded 10-minute participant traces are also displayed.

2. Related Work

2.1. Egocentric Datasets & Benchmarks

As shown in Table 1, early egocentric vision research [16–23] was established through foundational datasets like ADL [24], CharadesEgo [25], and EGTEA Gaze+ [26], though these were limited in scale. The field advanced significantly with larger-scale datasets such as EPIC-KITCHENS [4] and Ego4D [5], which broadened the scope to general daily tasks and established comprehensive benchmarks. Specialized datasets emerged to address specific challenges: EgoProceL [27] and IndustReal [28] for procedure learning, HoloAssist [29] for collaborative tasks, and EgoExo4D [8] and EgoExoLearn [9] for multiview understanding through integrated egocentric and exocentric perspectives. Recent benchmarks (shown in Table 2) built on Ego4D [5] and EPIC-KITCHENS [4] have advanced various aspects of first-person vision [30–33], including temporal understanding in EgoSchema [11] and planning in EgoPlan-Bench [34]. Recent advances in long-term egocentric video understanding have emerged with EgoMemoria [14] and HourVideo [15], yet multipersonal social dynamics and over-day habit patterns remain largely unexplored. EgoLife addresses this gap with a week-long, multiperson dataset that supports the analysis of prolonged behavioral patterns and complex social interactions, complemented by multimodal sensing, multiview perspectives, and detailed annotations.

2.2. Long-Context Video Language Models

Video-language models have progressed from classic video features extraction [35–41] to pretraining approaches [42–47] with enhanced capabilities, and currently to models designed to follow instructions [48–59]. More recent models [54, 57–68] and benchmarks [69–73] have focused on handling long-duration content, often spanning several hours, with solutions typically relying on video compression [55, 57, 61, 65, 74, 75] or extending model context length [60, 66, 67, 75, 76]. The EgoLife project pushes the boundary to week-long video content, potentially inspiring innovative approaches beyond conventional methods. For egocentric video-language models, while some models address egocentric content [30, 77–86] and attempt to handle longer video sequences [14, 15, 87–89], processing ultra-long egocentric footage remains an unexplored frontier.

3. The EgoLife Dataset & Benchmark

3.1. Data Collection

Overview The EgoLife dataset was collected over a seven-day period with six volunteers residing in a custom-designed environment, called the *EgoHouse* (shown in Figure 1). Each participant wore Meta Aria glasses [1] and captured multimodal egocentric videos. To enhance the dataset with third-person perspectives, 15 strategically placed GoPro cameras recorded the participants' activities from multiple angles. Additionally, millimeter-wave radars provided spatial and motion data, supporting synchronized, comprehensive multimodal analysis of daily events and interactions.

EgoLife Activities During the week, participants were asked to organize an Earth Day party on the second-to-last day. To prepare, they held meetings and discussions, rehearsed performances (such as music and dance), practiced and shared cooking skills, and decorated the house to align with the Earth Day theme. Activities extended beyond the house, as participants went shopping and sightseeing, with recording permission obtained in locations like shopping malls. Figure 3 shows the activity timeline for the week, and a detailed diary of the EgoLife week is in Appendix E.

Maintaining Informative and Coherent Capture We ensure that each pair of smart glasses records a minimum of six hours per day during participants' waking hours. To achieve this, the primary investigators actively monitor participants and provide gentle prompts to encourage engagement in meaningful activities when prolonged passive behavior, such as lying down and watching TikTok, is observed. Due to storage limitations, recordings are structured into three-hour segments. To maintain data continuity, the glasses are collected every three hours for data upload and storage clearance, a process that takes approximately one hour. During this period, participants are instructed to remain in their rooms and limit their activities to resting or non-essential tasks to prevent logic disruptions in the recorded footage.

Language The primary language of the EgoLife dataset is Chinese [1]. All the annotations (transcripts, captions, QAs) are primarily in Chinese and translated into English.

3.2. Data Cleaning

A rigorous data cleaning process was implemented to ensure synchronization, participant privacy, and readiness for annotation and data release, as illustrated in Figure 4.

3.3. Transcript Annotations

We started transcript annotation after synchronizing all the egocentric videos, merging audio tracks from six participants into one, and applying speech recognition [90] to generate initial timestamped transcripts. Using an open-source diarization algorithm [91], we differentiated the speakers and produced a preliminary transcript with overlapping conversations. This 50-hour transcript was then reviewed for accuracy. Afterward, we split the audio into six tracks, one for each participant. Reviewers refined each track, keeping only the speech audible to each participant, resulting in a final transcript accurately indicating who spoke each line.

[1] A one-day recording session with predominantly English speaking has also been conducted recently. More details are in Appendix.

Table 1. **Related Work for EgoLife Dataset - Overview of Egocentric Datasets.** For Modality, denotes video, denotes gaze, denotes IMU, denotes 3D scans. The EgoLife dataset stands out for its ultra-long egocentric footage and rich interpersonal interactions.

Benchmark	Domain	Modality	#Captions	Size (hrs)	#Clips	Dur./Clip	Multiview	Interpersonal Dynamics
EPIC-KITCHENS [4]	Kitchen		20K+	100	700	8.5 min	✗	✗
Ego4D [5]	Daily Activities		3.85M	3,670	9,645	22.8 min	✗	✗
EgoExo4D [8]	Skilled Activities		500K+	1,286	5,035	1 to 42 min	✓	✗
EgoExoLearn [9]	Task Execution		-	120	432	13.4 min	✓	✗
EgoPet [10]	Animal Actions		-	84	6,646	45.5 sec	✗	✗
EgoLife	Daily Life		400K+	266	6	44.3 h	✓	✓

Table 2. **Related Work for EgoLifeQA Benchmark.** The EgoLifeQA dataset is distinguished by its ultra-long video footage and certificate length, facilitating novel tasks such as habit discovery and relational interaction pattern analysis (see Figure 5 for details). **Note on Dur./Clip:** A clip is defined as a session with narrative continuity. For the EgoLife dataset, this value is derived from 266 hours of retained footage distributed across six participants.

Dataset	Source	#QAs	Size (hrs)	#Clips	Dur./Clip	Certificate Length [11]	
						Below 2h	Over 2h
EgoSchema [11]	Ego4D	5,063	250	5,063	3 min	5,063	0
EgoPlan-Bench [12]	Ego4D & EpicKitchen	4,939	-	4,939	-	4,939	0
EgoThink [13]	Ego4D	700	-	595	-	700	0
EgoMemoria [14]	Ego4D	7,026	-	629	30 s to 1 h	7,026	0
HourVideo [15]	Ego4D	12,976	381	500	20 min to 2 h	12,976	0
EgoLifeQA	**EgoLife**	3,000	266	6	44.3 h	997	2,003

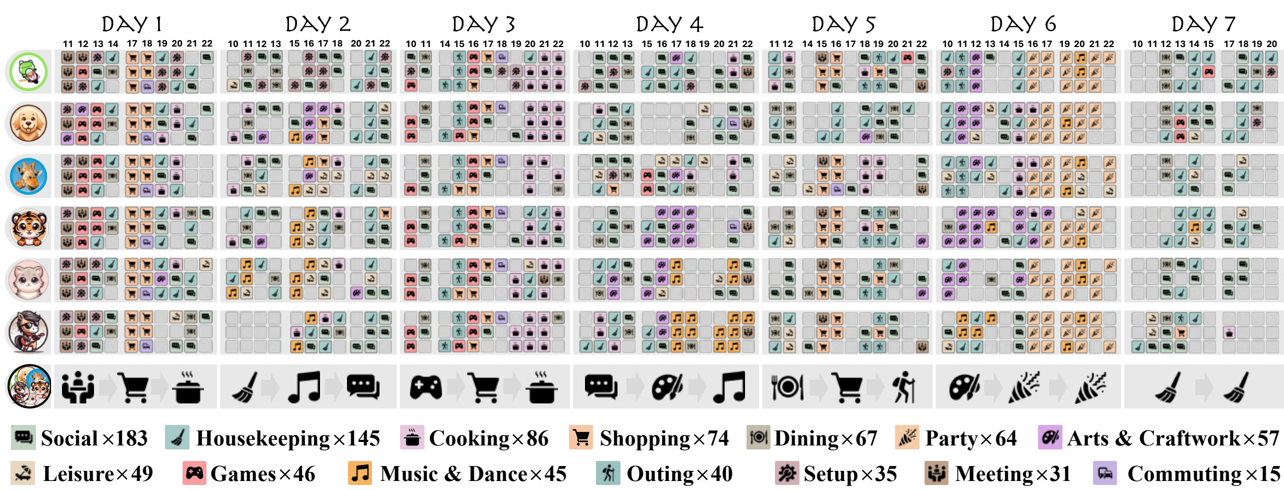


Figure 3. **The Activity Timeline of the EgoLife Dataset.** It visualizes the activity timeline of six participants over one week. Each block represents a 20-minute interval, color-coded and marked with icons for different activities. The legend shows 14 activity categories with their total occurrence counts. The categorization is automatically performed using GPT-4o on visual-audio captions with timestamps.

3.4. Caption Annotations

The captioning tool is a video editing software with dubbing functions [92]. We split all the videos into 5-minute clips, which were slowed to 0.8× speed, allowing annotators to provide continuous, detailed narrations by talking without pauses for high information density. Narration covered all actions, interactions, and notable environmental details. When no specific action was occurring, annotators described the participant's focus and prominent features in the surroundings. The narration was converted to text via a transcription tool, then reviewed and corrected for a synchronized, time-aligned textual description for each video segment.

The initial annotations, or "narrations," consisted of 361K brief, subtitle-like phrases, averaging 2.65 seconds per narration. Using GPT-4o-mini, we merged related phrases into 25K "merged captions," forming coherent sentences aligned with specific video segments. These captions were then expanded by pairing them with representative frames (sampled at 1 FPS) and corresponding transcripts, summarized by GPT-4o. This process transformed the "merged captions" into "visual-audio captions," which are enriched with both visual and speech context and verified by human annotators

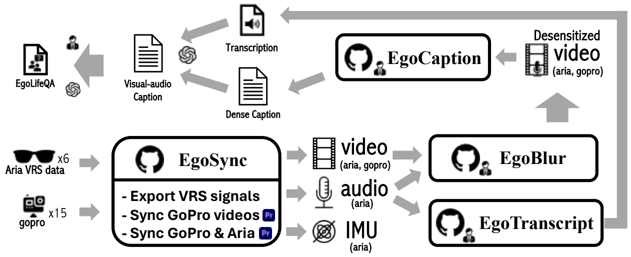


Figure 4. **The Overview of Data Process Pipeline.** The pipeline synchronizes multi-source data (video, audio, IMU) from Aria glasses and GoPro cameras using `EgoSync` codebase, processes them through privacy protection (`EgoBlur`), dense captioning (`EgoCaption`), and transcription (`EgoTranscript`) modules, ultimately feeding into the **EgoLifeQA** system.

(see Figure 1 for an example). These captions serve two main purposes: training EgoGPT and automatically generating QA candidates for the next section.

3.5. EgoLifeQA Annotations

For QA annotation, we designed five types of questions to assess the capabilities of a long-term life assistant:

- **EntityLog**: Tests long-term memory focused on object details like their last use, location, price, and more.
- **EventRecall**: Asks about past events and recalls details from the last time critical tasks were performed.
- **HabitInsight**: Focuses on personal habit patterns.
- **RelationMap**: Finds interpersonal interactions. This evaluates the performance of person identification.
- **TaskMaster**: Involves task assignment based on prior actions (e.g., reminding to buy a pen when the ink is low).

Examples of each question type can be found in Figure 5.

We crafted prompts for each type and fed "visual-audio captions" into GPT-4o in batches, generating around 100K timestamped questions per participant. These AI-generated questions were provided to annotators as SRT files, allowing them to view each question in sync with the relevant video segment. Rather than serving as final annotations, these questions acted as a filtering and inspiration tool for annotators, helping them identify valuable instances. Only questions requiring information from at least five minutes prior were retained, with a preference for those demanding longer dependencies and strong real-world relevance. This streamlined process enabled the efficient creation of a high-quality QA dataset tailored to long-context reasoning and practical real-world tasks.

After a rigorous selection and refinement process, we filtered the 100K QA candidates down to 1K high-quality questions per participant—less than 1% of the original pool—for further meticulous revision. This final round of curation resulted in a carefully crafted set of 500 QA per participant. Annotators also generated distractors for multiple-choice questions, formally establishing EgoLifeQA as a benchmark

Table 3. **Dataset Composition of EgoIT-99K.** We curated 9 classic egocentric video datasets and leveraged their annotations to generate captioning and QA instruction-tuning data for fine-tuning EgoGPT, building on the LLaVA-OneVision base model [55]. **#AV** means the number of videos with audio used for training. QAs include multiple types - **VC**: Video Captioning, **AVC**: Audio-Video Captioning, **MCQ**: Multiple Choice Questions, **MRC**: Multi-Round Questions, **IQA**: Image Question-Answering.

Dataset	Duration	#Videos (#AV)	#QA	QA Type
Ego4D [5]	3.34h	523 (458)	1.41K	VC, AVC, MCQ, MRC
Charades-Ego [25]	5.04h	591 (228)	18.46K	VC, AVC, MRC
HoloAssist [29]	9.17h	121	33.96K	VC, MCQ, MRC, IQA
EGTEA Gaze+ [26]	3.01h	16	11.20K	VC, MCQ, MRC, IQA
IndustReal [28]	2.96h	44	11.58K	VC, MCQ, MRC, IQA
EgoTaskQA [93]	8.72h	172	3.59K	VC, MCQ, MRC
EgoProceL [27]	3.11h	18	5.90K	VC, MCQ, MRC, IQA
Epic-Kitchens [4]	4.15h	36	10.15K	VC, MCQ, MRC, IQA
ADL [24]	3.66h	8	3.23K	VC, MCQ, MRC, IQA
Total	**43.16h**	**1529 (686)**	**99.48K**	

Table 4. **Performance of EgoGPT.** The table compares EgoGPT with state-of-the-art commercial and open-source models on existing egocentric benchmarks.

Model	#Param	#Frames	EgoSchema	EgoPlan	EgoThink
GPT-4v [94]	-	32	56.6	38.0	65.5
Gemini-1.5-Pro [95]	-	32	72.2	31.3	62.4
GPT-4o [96]	-	32	72.2	32.8	65.5
LLaVA-Next-Video [97]	7B	32	49.7	29.0	40.6
LongVA [98]	7B	32	44.1	29.9	48.3
IXC-2.5 [99]	7B	32	54.6	29.4	56.0
InternVideo2 [100]	8B	32	55.2	27.5	43.9
Qwen2-VL [101]	7B	32	66.7	34.3	59.3
Oryx [57]	7B	32	56.0	33.2	53.1
LLaVA-OV [55]	7B	32	60.1	30.7	54.2
LLaVA-Videos [102]	7B	32	57.3	33.6	56.4
EgoGPT (EgoIT)	7B	32	73.2	32.4	61.7
EgoGPT (EgoIT+EgoLifeD1)	7B	32	75.4	33.4	61.4

for multiple-choice question answering. Additionally, they annotated whether audio was required to answer the question and specified the look-back time (certification length) necessary for retrieving the correct answer. Statistical details are presented in Figure 6.

4. EgoButler: Agentic Egocentric Life Assistant

EgoButler is designed to tackle complex tasks presented by the EgoLifeQA. It comprises two core subsystems: **EgoGPT** (System-I) for clip-level omni-modal understanding and **EgoRAG** (System-II) for long-context question answering. The pipeline is illustrated in Figure 7.

4.1. System-I: EgoGPT for Clip Understanding

EgoGPT has two main functions in EgoButler. First, it performs continuous video captioning: processing each 30-second clip to generate captions using both visual and audio inputs. This multimodal captioning provides immediate understanding and valuable context for EgoRAG retrieval tasks. Second, EgoGPT assists with question-answering by utilizing retrieved clues from EgoRAG.

To better align with the egocentric video domain and incorporate audio understanding, we introduce EgoIT-99K,

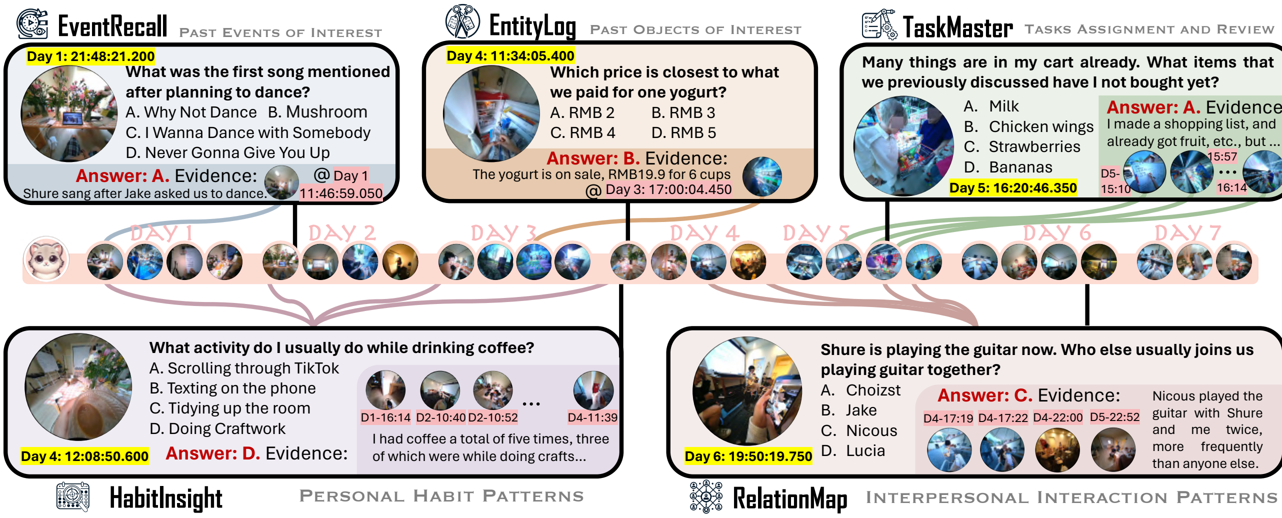


Figure 5. **Question Types and Examples in the EgoLifeQA Benchmark.** We design five types of questions to evaluate egocentric assistants' capabilities in entity logging, event recall, task tracking, and human-centric problems (habit analysis and relationship understanding). Each example includes a multiple-choice Q&A with supporting evidence from timestamps at least 5 minutes prior to the question. Black vertical lines indicate question timestamps, while colored curved lines connect to relevant evidence timestamps.

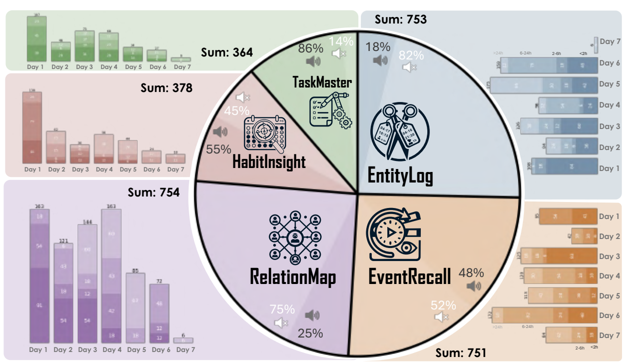


Figure 6. **Statistics of EgoLifeQA.** We gathered 500 long-context QAs per participant, totaling 3K QAs. The sum of QAs for each question type is reported. In the pie chart, darker segments indicate the proportion of questions requiring audio. The bar chart presents the daily count of QAs per question type, with brightness levels reflecting 4-level certification length [11] (from <2h to >24h).

a diverse and representative egocentric video dataset (detailed in Table 3) with QA pairs derived from video annotations using prompts tailored to actions, objects, and events (see Appendix F). This dataset is used to fine-tune EgoGPT on LLaVA-OneVision [55], incorporating videos with audio as training data. Since LLaVA-OneVision is built on Qwen2, we develop an audio branch similar to Ola [58], encoding audio with Whisper Large v3 [90] and training an audio projection module on LibriSpeech [103]. Starting from the audio projection module upon LLaVA-OneVision, we use EgoIT-99K for final stage finetuning. For personalization, we fine-tune EgoGPT on EgoLife Day-1's video, enabling identity-aware questioning in EgoLifeQA. We define EgoGPT (EgoIT-99K+D1) as the personalized version and EgoGPT (EgoIT-99K) as the non-personalized baseline.

4.2. System-II: EgoRAG for Long-Context Q&A

To address long-horizon, long-context scenarios, EgoRAG—a retrieval-augmented generation (RAG) system—enhances memory and query capabilities, enabling personalized and long-term comprehension. It employs a two-stage approach:

Memory Bank Construction In the first stage, EgoRAG integrates with EgoGPT to extract video clip captions and store them in a structured memory module, ensuring efficient retrieval of time-stamped contextual information. Captions are continuously generated by EgoGPT and summarized at hourly and daily levels by a language model, forming a multi-level memory bank for scalable retrieval. The memory bank M consists of:

$$M = \{(c_i, d_i, t_i)\}_{i=1}^{N} \tag{1}$$

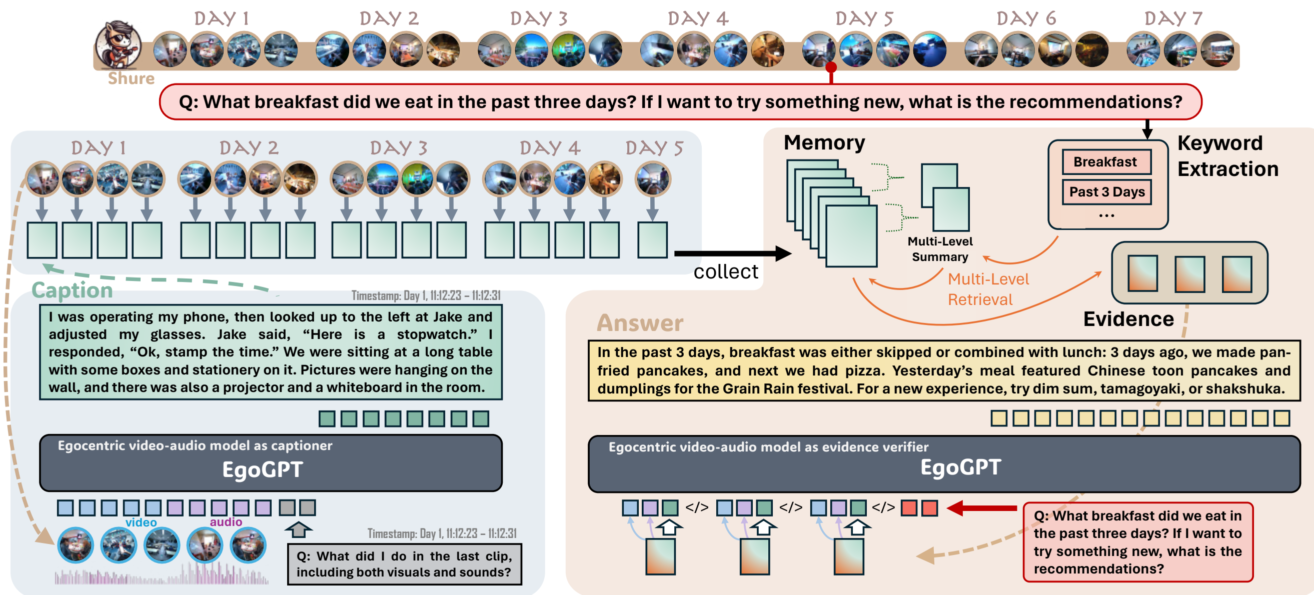
where c_i represents clip features, d_i textual descriptions, and t_i timestamped summaries (hourly, daily).

Content Retrieval and Response Generation When a question is posed, EgoRAG hypothesizes the relevant time window by first retrieving higher-level summaries t_i and refining the search from day to hour. Within the selected window, fine-grained retrieval is performed using a relevance-based scoring function:

$$s_i = \text{Similarity}(q, c_i) + \lambda \text{Similarity}(q, d_i), \tag{2}$$

where λ balances visual and textual relevance. The top-k most relevant clips are selected:

$$R = \text{TopK}(\{(c_i, d_i, s_i)\}_{i=1}^{N}). \tag{3}$$



(a) Captioning Stage with EgoGPT **(b) Question Answering Stage with EgoRAG**

Figure 7. **The EgoBulter Architecture.** The system comprises (a) a Captioning Stage powered by EgoGPT for dense visual-audio understanding of egocentric clips, and (b) a Question Answering Stage utilizing EgoRAG for memory retrieval and response generation. The example demonstrates temporal reasoning across multiple days, with keyword extraction, evidence retrieval, and context-aware answer generation for a breakfast-related query.

Table 5. **Performance comparison of EgoGPT with state-of-the-art models on EgoLifeQA benchmarks.** For a fair comparison on EgoLifeQA, EgoGPT was replaced with the corresponding models in the EgoButler pipeline to evaluate their performance under the same conditions. Models that provide captions for EgoLifeQA use 1 FPS for video sampling.

Model	#Frames	Audio	Identity	EgoLifeQA					
				EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Average
Gemini-1.5-Pro [95]	-	✓	✗	36.0	37.3	45.9	30.4	34.9	36.9
GPT-4o [96]	1 FPS	✗	✗	34.4	42.1	29.5	30.4	44.4	36.2
LLaVA-OV [55]	1 FPS	✗	✗	36.8	34.9	31.1	22.4	28.6	30.8
EgoGPT (EgoIT-99K)	1 FPS	✓	✗	35.2	36.5	27.9	29.6	36.5	33.1
EgoGPT (EgoIT-99K+D1)	1 FPS	✓	✓	39.2	36.5	31.1	33.6	39.7	36.0

The retrieved content is then fed into a language model (EgoGPT, GPT-4o, etc.) to generate an informed response:

$$r = \text{EgoGPT/GPT}(q, R). \tag{4}$$

This hierarchical retrieval strategy ensures that responses are both contextually relevant and computationally efficient.

4.3. Integration and Synergy in EgoButler

Together, EgoGPT and EgoRAG form the EgoButler system, combining efficient video interpretation with long-context memory. EgoGPT continuously gathers personalized egocentric data, while EgoRAG retrieves and delivers relevant clues, enabling accurate and context-aware responses.

5. Experiments

Implementation Details We evaluate EgoGPT (7B) on three egocentric datasets: EgoSchema [11], EgoPlan-Bench [12], and EgoThink [13], using 32 video frames per clip where applicable for fair comparison. For EgoLifeQA, we conduct a quick evaluation on Jake's 500 QA in this version. To compare different models, we integrate them into the EgoButler framework as captioners, replacing EgoGPT while collaborating with EgoRAG for QA tasks. The final response is universally generated by GPT-4o for fair evaluation (see Eq. 4). EgoRAG follows a simple retrieval pipeline: text-based similarity retrieval (setting $\lambda = 0$ in Eq. 2) selects the top 3 most relevant 30-second clips as input to EgoGPT and its alternatives. Re-querying is performed using GPT-4o-mini with pre-stored results to ensure fairness.

Table 6. **Effectiveness of EgoRAG.** Integrating EgoRAG significantly enhances video-language models' performance in long-context question answering, especially for questions requiring longer certification lengths. For comparison, we evaluate Gemini-1.5-Pro and EgoGPT on a half-hour video segment, limiting their answers to this timeframe.

Model	Certificate Length			
	$< 2h$	$2h - 6h$	$6h - 24h$	$> 24h$
Gemini-1.5-Pro	27.9	14.8	25.0	18.4
EgoGPT	28.2	29.1	26.8	25.0
EgoGPT+EgoRAG	27.2	35.7	38.9	35.4

Main Results of EgoGPT Table 4 presents a performance

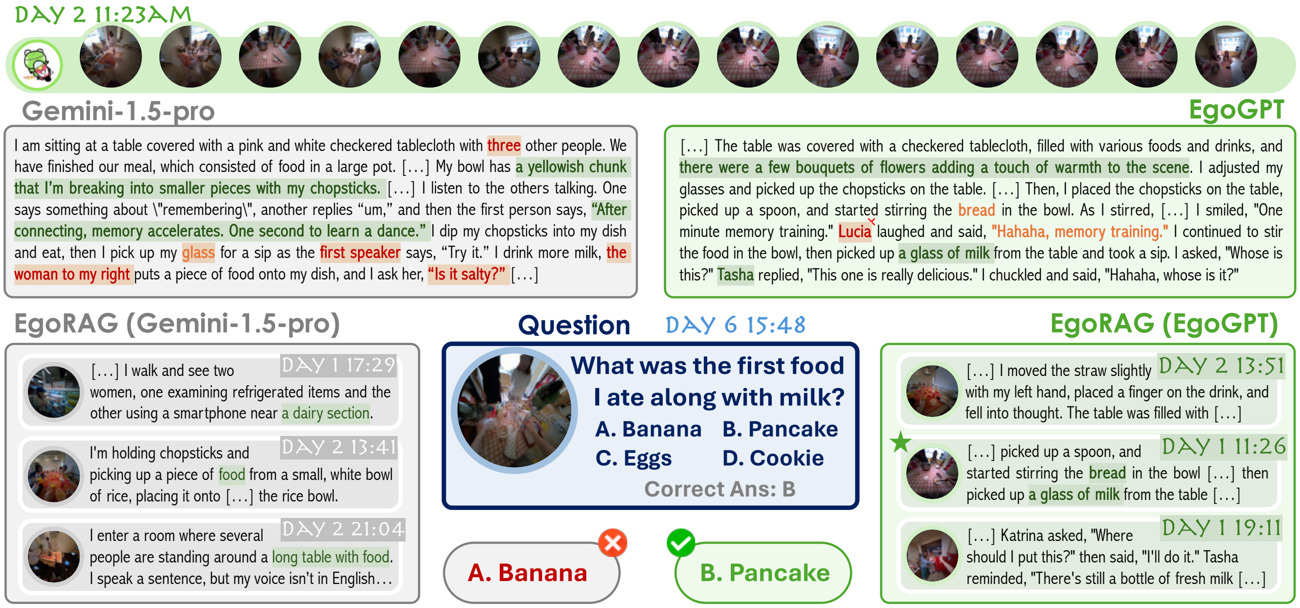


Figure 8. **Qualitative Comparison of EgoGPT and Gemini-1.5-Pro under the EgoButler Framework.** The top section compares captions from two models on a 30-second clip: EgoGPT excels in personalization and hallucinates less on the egocentric videos. The bottom section features a question that is answered by the clip, showcasing EgoRAG's skill in pinpointing relevant time slots and key clues.

Table 7. **Ablation Study on EgoGPT.** We construct different EgoRAG memory banks using generated captions from EgoGPT variants. The first three rows use captions from human annotations as a reference. All response generation models utilize EgoGPT (EgoIT-99K+D1) to ensure fair comparison. The result indicates how caption quality affects of EgoBulter performance.

Caption Source	Visual	Audio	Dataset	Avg.
Narration	✓	✓	-	31.5
Transcript	✗	✓	-	29.6
Visual-Audio Caption	✓	✓	-	45.5
EgoGPT (Audio Only)	✗	✓	EgoIT-99K	27.2
EgoGPT (Audio Only)	✗	✓	EgoIT-99K+D1	28.1
EgoGPT (Visual Only)	✓	✗	EgoIT-99K	31.2
EgoGPT (Visual Only)	✓	✗	EgoIT-99K+D1	33.6
EgoGPT (Visual+Audio)	✓	✓	EgoIT-99K	33.1
EgoGPT (Visual+Audio)	✓	✓	EgoIT-99K+D1	36.0

comparison of EgoGPT with state-of-the-art commercial and open-source models on egocentric benchmarks. Powered by the EgoIT-99K dataset, EgoGPT demonstrates strong performance across these benchmarks, with EgoGPT (EgoIT-99K+D1) achieving the highest average score. For Table 5, EgoGPT's ability to recognize individuals and integrate omni-modal information effectively distinguishes it from general-purpose commercial models like GPT-4o and Gemini-1.5-Pro, which lack personalized adaptation. However, while EgoGPT shows notable advantages in certain areas, particularly in RelationMap and omni-modal integration, the task remains inherently challenging, and there is still a large room for improvement.

The Effects of EgoRAG Table 6 highlights the impact of EgoRAG on long-context question answering. Models like Gemini-1.5-Pro and EgoGPT cannot process ultra-long videos exceeding 40 hours. To handle this, we split the videos into 30-minute segments and posed questions directly within each segment. This allows the models to answer without requiring EgoRAG. However, this segmentation approach often results in hallucinations and incorrect answers due to the lack of global context, especially for questions that require clues from other segments. EgoRAG mitigates these issues by retrieving relevant evidence across segments, significantly improving accuracy. For queries spanning over 24 hours, EgoGPT+EgoRAG achieves a score of 35.4, outperforming both EgoGPT and Gemini-1.5-Pro, demonstrating the critical role of long-term retrieval.

Analysis of EgoGPT Variants Table 7 shows that human captions yield the highest scores, emphasizing the importance of quality captions for better retrieval. Audio-only models perform weakest, while visual-only models perform better, suggesting audio alone isn't enough for EgoLifeQA. Combining visual and audio inputs achieves the best performance, with improvements from adding EgoLife Day-1 captions.

Qualitative Results Figure 8 shows EgoGPT excels in personalization and contextually relevant captions but struggles with speech comprehension, especially emotions and laughter. It also overfits Day-1 data, misidentifying individuals. EgoRAG retrieves long-context evidence but lacks multi-step reasoning, failing when key information is missing. Future improvements should focus on speech understanding, personalization, and advanced retrieval.

Acknowledgement

This study is supported by the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative. We would like to sincerely thank Meta Aria for their generous sponsorship, which has greatly supported the success of this project.

References

[1] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561, 2023. 1, 2, 3

[2] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. Neurocomputing, 472:175–197, 2022. 2

[3] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. International Journal of Computer Vision, pages 1–57, 2024. 2

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, et al. The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11):4125–4141, 2021. 2, 3, 4, 5, 22, 24, 25

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18995–19012, June 2022. 2, 3, 4, 5, 22, 24, 25

[6] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19383–19400, 2024. 2

[7] Fahad Jibrin Abdu, Yixiong Zhang, Maozhong Fu, Yuhan Li, and Zhenmiao Deng. Application of deep learning on millimeter-wave radar signals: A review. Sensors, 21(6):1951, 2021. 2

[8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19383–19400, June 2024. 3, 4, 26

[9] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Dong Lu, Yali Wang, 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, 2024. 3, 4, 26

[10] Amir Bar, Arya Bakhtiar, Danny Tran, Antonio Loquercio, Jathushan Rajasegaran, Yann LeCun, Amir Globerson, and Trevor Darrell. Egopet: Egomotion and interaction data from an animal's perspective. arXiv preprint arXiv:2404.09991, 2024. 4

[11] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. 3, 4, 6, 7, 25

[12] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. arXiv preprint arXiv:2312.06722, 2023. 4, 7

[13] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14291–14302, 2024. 4, 7

[14] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms. arXiv preprint arXiv:2410.07177, 2024. 3, 4

[15] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding, 2024. 3, 4

[16] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 3, 26

[17] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1226–1233. IEEE, 2012. 23

[18] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In 2012 IEEE conference on computer vision and pattern recognition, pages 1346–1353. IEEE, 2012. 23

[19] Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2730–2737, 2013. 23

[20] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE international conference on computer vision, pages 1949–1957, 2015. 25

[21] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocen-

tric rgb-d sensor. In Proceedings of the IEEE international conference on computer vision, pages 1154–1163, 2017.

[22] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. IEEE transactions on pattern analysis and machine intelligence, 41(7):1720–1733, 2018. 24

[23] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. arXiv preprint arXiv:2408.00754, 2024. 3

[24] Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. Youhome system and dataset: Making your home know you better. IEEE International Symposium on Smart Electronic Systems (IEEE - iSES), 2022. 3, 5, 22

[25] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos, 2018. 3, 5, 22, 26

[26] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In European Conference on Computer Vision (ECCV), 2018. 3, 5, 22, 24

[27] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In European Conference on Computer Vision (ECCV), 2022. 3, 5, 22, 25

[28] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, et al. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 4365–4374, 2024. 3, 5, 22, 25

[29] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, 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, pages 20270–20281, 2023. 3, 5, 22, 25

[30] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. Advances in Neural Information Processing Systems, 35:7575–7586, 2022. 3, 25

[31] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. Advances in Neural Information Processing Systems, 35:13745–13758, 2022. 24

[32] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 19935–19947, 2022. 24

[33] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22836–22845, 2023. 3, 25

[34] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13299–13308, 2024. 3

[35] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1657–1666, 2017. 3

[36] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In Proceedings of the IEEE international conference on computer vision, pages 4193–4202, 2017.

[37] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017.

[38] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6504–6512, 2017.

[39] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6576–6585, 2018.

[40] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8739–8748, 2018.

[41] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1999–2007, 2019. 3

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7464–7473, 2019. 3

[43] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8746–8755, 2020.

[44] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. arXiv preprint arXiv:2005.05402, 2020.

[45] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9972–9981, 2020.

[46] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681, 2021.

[47] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022. 3

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 3

[49] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 543–553, Singapore, December 2023. Association for Computational Linguistics.

[50] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[51] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023.

[52] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023.

[53] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[54] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024. 3

[55] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024. 3, 5, 6, 7

[56] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13700–13710, 2024.

[57] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. arXiv preprint arXiv:2409.12961, 2024. 3, 5

[58] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. arXiv preprint arXiv:2502.04328, 2025. 6

[59] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. arXiv preprint arXiv:2411.14432, 2024. 3

[60] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. arXiv preprint arXiv:2406.16852, 2024. 3

[61] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv:2410.17434, 2024. 3

[62] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, 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, pages 18221–18232, 2024.

[63] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14313–14323, 2024.

[64] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840, 2024.

[65] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. arXiv preprint arXiv:2408.15542, 2024. 3

[66] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In European Conference on Computer Vision, pages 453–470. Springer, 2025. 3

[67] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188, 2024. 3

[68] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. arXiv preprint arXiv:2403.12966, 2024. 3

[69] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. arXiv preprint arXiv:2312.04817, 2023. 3

[70] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. arXiv preprint arXiv:2407.15754, 2024.

[71] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035, 2024.

[72] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024.

[73] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813, 2024. 3

[74] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In European Conference on Computer Vision, pages 323–340. Springer, 2025. 3

[75] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26689–26699, 2024. 3

[76] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. arXiv preprint arXiv:2402.08268, 2024. 3

[77] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6943–6953, 2021. 3

[78] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. arXiv preprint arXiv:2210.14395, 2022.

[79] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23066–23078, 2023.

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

[81] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3307–3317, October 2023.

[82] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. Advances in Neural Information Processing Systems, 36:53688–53710, 2023.

[83] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5250–5261, 2023.

[84] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems, 36, 2024.

[85] Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, and Giuseppe Averta. A backpack full of skills: Egocentric video understanding with diverse task perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18275–18285, June 2024.

[86] Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torresani, and Effrosyni Mavroudi. Learning to segment referred objects from narrated egocentric videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14510–14520, June 2024. 3

[87] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Lalm: Long-term action anticipation with language models. arXiv preprint arXiv:2311.17944, 2023. 3

[88] Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbron, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 13667–13678, October 2023.

[89] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18198–18208, 2024. 3

[90] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In International conference on machine learning, pages 28492–28518. PMLR, 2023. 3, 6

[91] Mahmoud Ashraf. Whisper diarization: Speaker diarization using openai whisper. 2024. 3

[92] ByteDance. Capcut, 2024. Mobile application software. 4

[93] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. Advances in Neural Information Processing Systems, 35:3343–3360, 2022. 5, 22, 26

[94] OpenAI. Gpt-4v(ision) system card, 2023. 5

[95] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. 5, 7

[96] OpenAI. Gpt-4o system card, 2024. 5, 7

[97] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. 5

[98] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. arXiv preprint arXiv:2406.16852, 2024. 5

[99] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112, 2023. 5

[100] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. arXiv preprint arXiv:2403.15377, 2024. 5

[101] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759, 2024. 5

[102] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024. 5

[103] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 5206–5210. IEEE, 2015. 6

[104] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In CVPR 2011, pages 3241–3248. IEEE, 2011. 23

[105] Omid Aghazadeh, Josephine Sullivan, and Stefan Carlsson. Novelty detection from an ego-centric perspective. In CVPR 2011, pages 3297–3304. IEEE, 2011. 23

[106] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In 2012 IEEE conference on computer vision and pattern recognition, pages 2847–2854. IEEE, 2012. 23

[107] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In BMVC, volume 2, page 3. Citeseer, 2014. 23

[108] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2537–2544, 2014. 23

[109] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In Proceedings of the IEEE International Conference on Computer Vision, pages 4498–4506, 2015. 23

[110] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1868–1877, 2017. 23

[111] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4372–4381, 2017. 23

[112] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In 2017 IEEE International Conference on Image Processing (ICIP), pages 3410–3414. IEEE, 2017. 23, 25

[113] Michel Silva, Washington Ramos, Joao Ferreira, Felipe Chamone, Mario Campos, and Erickson R Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2383–2392, 2018. 23

[114] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 637–652, 2018. 23

[115] Emiliano Spera, Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Egocentric shopping cart localization. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 2277–2282. IEEE, 2018. 23

[116] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019. 24

[117] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. Ego-ch: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. Pattern Recognition Letters, 131:150–157, 2020. 24

[118] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 24

[119] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9890–9900, 2020. 24

[120] Haonan Qiu, Pan He, Shuchun Liu, Weiyuan Shao, Feiyun Zhang, Jiajun Wang, Liang He, and Feng Wang. Ego-deliver: A large-scale dataset for egocentric video analysis. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1847–1855, 2021. 24

[121] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. arXiv preprint arXiv:2211.12498, 2022. 24

[122] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21013–21022, 2022. 24

[123] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 20110–20120, 2023. 24

[124] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 20133–20143, 2023. 24

[125] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition. arXiv preprint arXiv:2304.05088, 2023. 24

[126] Xueyi Wang. Egofalls: a visual-audio dataset and benchmark for fall detection using egocentric cameras. arXiv preprint arXiv:2309.04579, 2023. 24

[127] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2698–2710, 2021. 24

[128] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22910–22921, 2023. 24

[129] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic video scene graph generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18675–18685, 2023. 25

[130] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epictent: An egocentric video dataset for camping tent assembly. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019. 25

[131] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1569–1578, 2021. 25

[132] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21096–21106, 2022. 25

[133] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In European Conference on Computer Vision, pages 485–501. Springer, 2022. 25

[134] Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, and Giovanni Maria Farinella. Enigma-51: Towards a fine-grained understanding of human-object interactions in industrial scenarios. arXiv preprint arXiv:2309.14809, 2023. 25

[135] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. arXiv preprint arXiv:2311.16444, 2023. 25

[136] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018. 25

[137] Xiaofeng Ren and Matthai Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 1–8. IEEE, 2009. 25

[138] Cheng Li and Kris M Kitani. Pixel-level hand detection in ego-centric videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3570–3577, 2013. 25

[139] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia, 20(5):1038–1050, 2018. 25

[140] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of International Conference on Computer Vision (ICCV), 2017. 25

[141] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 409–419, 2018. 25

[142] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10138–10148, 2021. 25

[143] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20971–20980. IEEE, 2022. 25

[144] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In European Conference on Computer Vision, pages 127–145. Springer, 2022. 25

[145] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12999–13008, 2023. 25

[146] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. arXiv preprint arXiv:2411.08380, 2024. 25

[147] Banerjee Prithviraj, Shkodrani Sindi, Moulonand Pierre, Hampali Shreyas, and Han Shangchen. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. arXiv preprint arXiv:2411.19167, 2024. 25

[148] Zhao Yiming, Kwon Taein, Streli Paul, Pollefeys Marc, and Holz Christian. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. arXiv preprint arXiv:2409.02224, 2024. 25

[149] Ashutosh Kumar, Nagarajan Tushar, Pavlakos Georgios, Kitani Kris, and Grauman Kristen. Expertaf: Expert actionable feedback from video. arXiv preprint arXiv:2408.00672, 2024. 25

[150] Fujii Ryo, Saito Hideo, and Kajita Hiroki. Egosurgery-tool: A dataset of surgical tool and hand detection from egocentric open surgery videos. arXiv preprint arXiv:2406.03095, 2024. 25

[151] Fujii Ryo, Hatano Masashi, Saito Hideo, and Kajita Hiroki. Egosurgery-phase: A dataset of surgical phase recognition from egocentric open surgery videos. arXiv preprint arXiv:2405.19644, 2024. 25

[152] Qiu Lu, Ge Yuying, Chen Yi, Ge Yixiao, Shan Ying, and Liu Xihui. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. arXiv preprint arXiv:2412.04447, 2024. 25

[153] Yuan Huaying, Ni Jian, Wang Yueze, Zhou Junjie, Liang Zhengyang, Liu Zheng, Cao Zhao, Dou Zhicheng, and Wen Ji-Rong. Momentseeker: A comprehensive benchmark and a strong baseline for moment retrieval within long videos. arXiv preprint arXiv:2502.12558, 2025. 25

[154] Zhang Wenyu, En Ng Wei, Ma Lixin, Wang Yuwen, Zhao Jungqi, Koenecke Allison, Li Boyang, and Wang Lu. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. arXiv preprint arXiv:2412.12693, 2024. 25

[155] Zhou Sheng, Xiao Junbin, Li Qingyun, Li Yicong, Yang Xun, Guo Dan, Wang Meng, Chua Tat-Seng, and Yao Angela. Egotextvqa: Towards egocentric scene-text aware video question answering. arXiv preprint arXiv:2502.07411, 2025. 25

[156] M. Waghmare Sagar, Wilber Kimberly, Hawkey Dave, Yang Xuan, Wilson Matthew, Debats Stephanie, Nuengsigkapian Cattalyya, Sharma Astuti, Pandikow Lars, Wang Huisheng, Adam Hartwig, and Sirotenko Mikhail. Sanpo: A scene understanding, accessibility and human navigation dataset. arXiv preprint arXiv:2309.12172, 2023. 25

[157] Plizzari Chiara, Goel Shubham, Perrett Toby, Chalk Jacob, Kanazawa Angjoo, and Damen Dima. Spatial cognition from egocentric video: Out of sight, not out of mind. arXiv preprint arXiv:2404.05072, 2024. 25

[158] Zhou Junjie, Shu Yan, Zhao Bo, Wu Boya, Liang Zhengyang, Xiao Shitao, Qin Minghao, Yang Xi, Xiong Yongping, Zhang Bo, Huang Tiejun, and Liu Zheng. Mlvu: Benchmarking multi-task long video understanding. arXiv preprint arXiv:2406.04264, 2024. 25

[159] Zhu Xilei, Duan Huiyu, Yang Liu, Zhu Yucheng, Min Xiongkuo, Zhai Guangtao, and Le Callet Patrick. Esvqa: Perceptual quality assessment of egocentric spatial videos. arXiv preprint arXiv:2412.20423, 2024. 25

[160] Darkhalil Ahmad, Guerrier Rhodri, W. Harley Adam, and Damen Dima. Egopoints: Advancing point tracking for egocentric videos. arXiv preprint arXiv:2412.04592, 2024. 25

[161] Qiu Heqian, Shi Zhaofeng, Wang Lanxiao, Xiong Huiyu, Li Xiang, and Li Hongliang. Egome: Follow me via egocentric view in real world. arXiv preprint arXiv:2501.19061, 2025. 25

[162] Nishimoto Tomohiro, Nishimura Taichi, Yamamoto Koki, Shirai Keisuke, Kameko Hirotaka, Haneji Yuto, Yoshida Tomoya, Kajimura Keiya, Cui Taiyu, Nishiwaki Chihiro, Daikoku Eriko, Okuda Natsuko, Ono Fumihito, and Mori Shinsuke. Biovl-qr: Egocentric biochemical vision-and-language dataset using micro qr codes. arXiv preprint arXiv:2404.03161, 2024. 25

[163] Kadambi Adesh and Zariffa José. Detecting activities of daily living in egocentric video to contextualize hand use at home in outpatient neurorehabilitation settings. arXiv preprint arXiv:2412.10846, 2024. 25

[164] Haneji Yuto, Nishimura Taichi, Kameko Hirotaka, Shirai Keisuke, Yoshida Tomoya, Kajimura Keiya, Yamamoto Koki, Cui Taiyu, Nishimoto Tomohiro, and Mori Shinsuke. Egooops: A dataset for mistake action detection from egocentric videos referring to procedural texts. arXiv preprint arXiv:2410.05343, 2024. 25

[165] Chen Lu, Wang Yizhou, Tang Shixiang, Ma Qianhong, He Tong, Ouyang Wanli, Zhou Xiaowei, Bao Hujun, and Peng Sida. Acquisition through my eyes and steps: A joint predictive agent model in egocentric worlds. arXiv preprint arXiv:2502.05857, 2025. 25

[166] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2629–2638, 2016. 26

[167] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In European Conference on Computer Vision, pages 767–786. Springer, 2020. 26

[168] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11184–11193, 2021. 26

[169] Mohamed Elfeki, Liqiang Wang, and Ali Borji. Multi-stream dynamic video summarization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 339–349, 2022. 26

[170] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people

from head-mounted devices. In European conference on computer vision, pages 180–200. Springer, 2022. 26