

EgoLife: Towards Egocentric Life Assistant



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 identity recognition, and 3) facilitating long-context question answering over extensive temporal information, we introduce **EgoButler**, 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.

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

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.

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 proce-

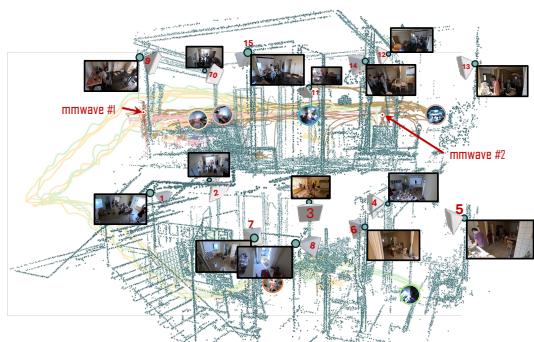


Figure 2. 3D reconstruction of the shared house using Aria MultiPS [1], showcasing the locations of 15 Exo cameras in the common area and 2 mmWave devices (highlighted in red) on the second floor. Color-coded 10-minute participant traces are also displayed.

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	6,000	266	6	44.3 h	997	2,003

ture 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 everyday 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

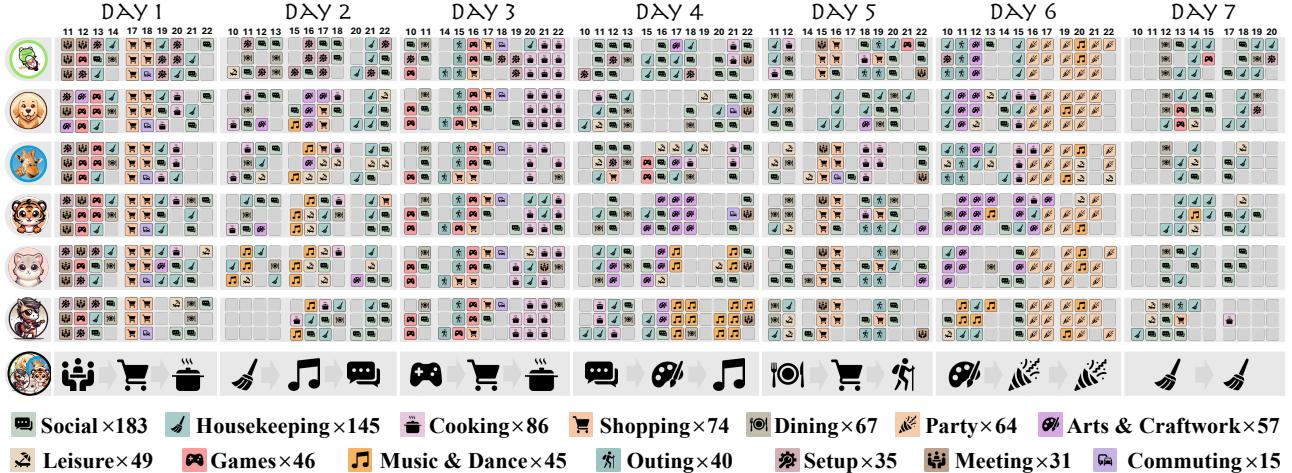


Figure 3. **The Activity Timeline of the EgoLife Dataset.** It visualizes the activity timeline of six participants over one week. Each block represents a 20-minute interval, color-coded and marked with icons for different activities. The legend shows 14 activity categories with their total occurrence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

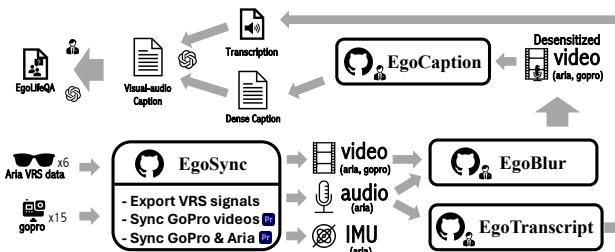


Figure 4. **The Overview of Data Process Pipeline.** The pipeline synchronizes multi-source data (video, audio, IMU) from Aria glasses and GoPro cameras using EgoSync codebase, processes them through privacy protection (EgoBlur), dense captioning (EgoCaption), and transcription (EgoTranscript) modules,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

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 Chinese¹. 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.

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

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.

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 $0.8\times$ 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

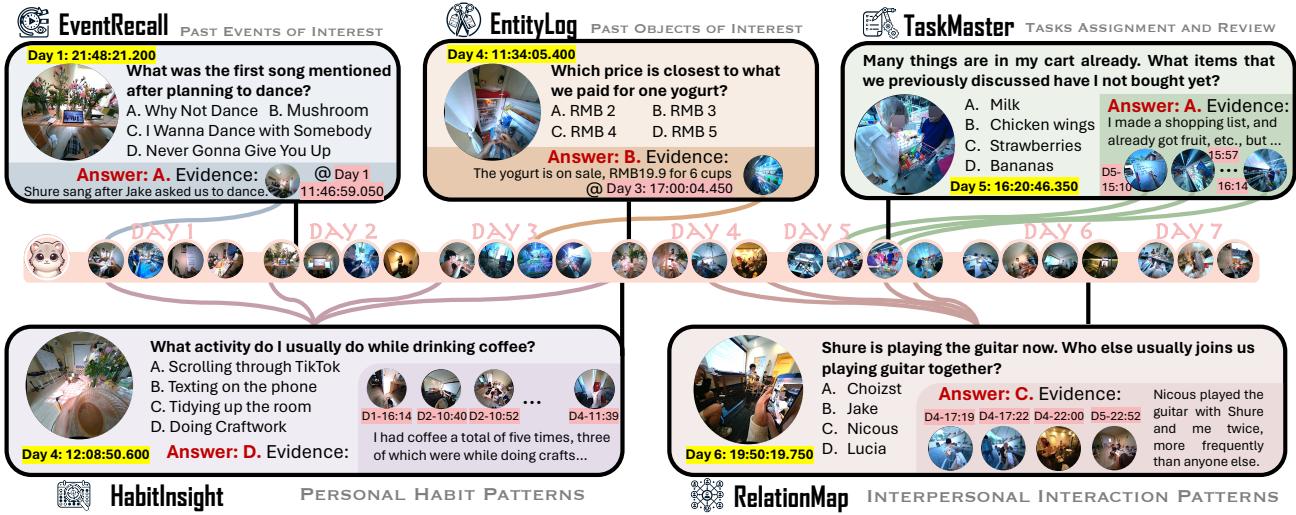


Figure 5. **Question Types and Examples in the EgoLifeQA Benchmark.** We design five types of questions to evaluate egocentric assistants’ capabilities in entity logging, event recall, task tracking, and human-centric problems (habit analysis and relationship understanding). Each example includes a multiple-choice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

(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 fil-

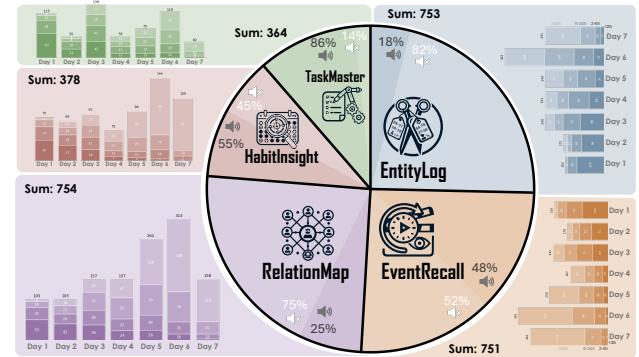


Figure 6. **Statistics of EgoLifeQA.** We gathered 500 long-context QAs per participant, totaling 3K QAs. The sum of QAs for each question type is reported. In the pie chart, darker segments indicate the proportion of questions requiring audio. The bar chart presents the daily count of QAs per question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

tered 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 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**

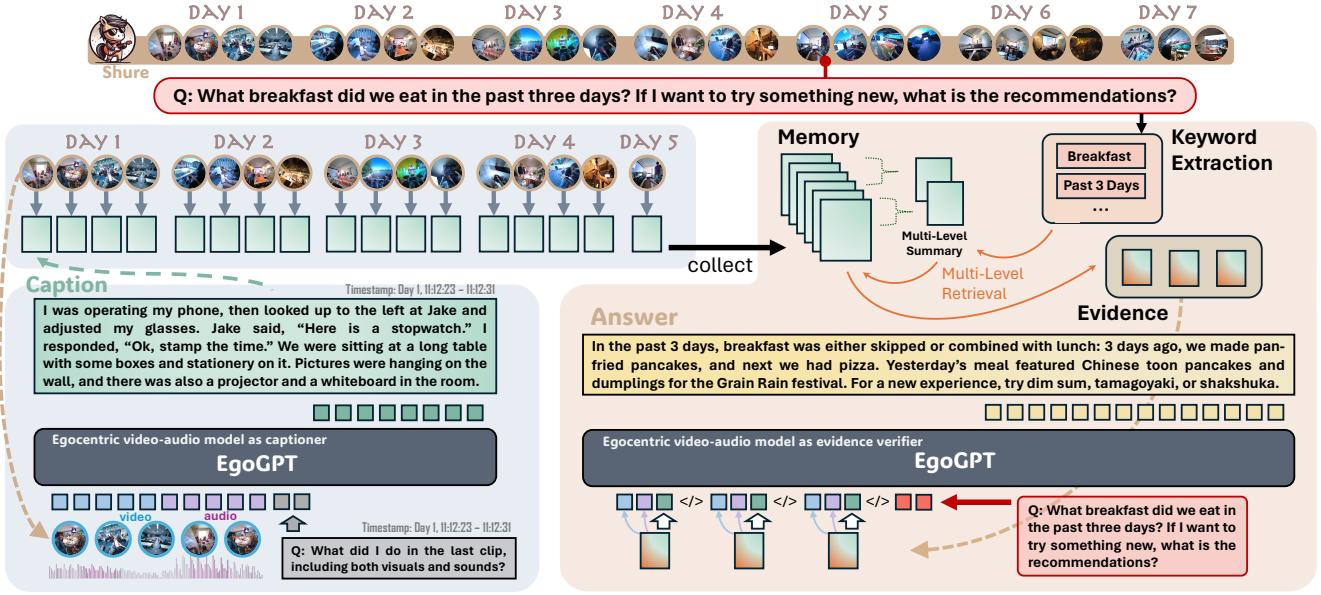


Figure 7. **The EgoButler Architecture.** The system comprises (a) a Captioning Stage powered by EgoGPT for dense visual-audio understanding of egocentric clips, and (b) a Question Answering Stage utilizing EgoRAG for memory retrieval and response generation. The example demonstrates temporal reasoning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

Table 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
EgoProcE [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	

(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, a diverse and representative egocentric video dataset (de-

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

tailed 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)

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	EntityLog	EgoLifeQA				
					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

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:

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

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:

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

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

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

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

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

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.

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

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.

Main Results of EgoGPT Table 4 presents a performance 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.

DAY 2 11:23AM

**Gemini-1.5-pro**

I am sitting at a table covered with a pink and white checkered tablecloth with **three** other people. We have finished our meal, which consisted of food in a large pot. [...] My bowl has a yellowish chunk **that I'm breaking into smaller pieces with my chopsticks.** [...] I listen to the others talking. One says something about 'remembering', another replies 'um,' and then the first person says, **"After connecting, memory accelerates. One second to learn a dance."** I dip my chopsticks into my dish and eat, then I pick up my **glass** for a sip as the **first speaker** says, 'Try it.' I drink more milk, **the woman to my right** puts a piece of food onto my dish, and I ask her, **"Is it salty?"** [...]

EgoGPT

[...] The table was covered with a checkered tablecloth, filled with various foods and drinks, and **there were a few bouquets of flowers adding a touch of warmth to the scene.** I adjusted my glasses and picked up the chopsticks on the table. [...] Then, I placed the chopsticks on the table, picked up a spoon, and started stirring the **bread** in the bowl. As I stirred, [...] I smiled, 'One minute memory training.' **Lucia** laughed and said, **"Hahaha, memory training."** I continued to stir the food in the bowl, then picked up a **glass of milk** from the table and took a sip. I asked, 'Whose is this?' **Tasha** replied, 'This one is really delicious.' I chuckled and said, 'Hahaha, whose is it?'

EgoRAG (Gemini-1.5-pro)

	[...] I walk and see two women, one examining refrigerated items and the other using a smartphone near a dairy section.	DAY 1 17:29
	I'm holding chopsticks and picking up a piece of food from a small, white bowl of rice, placing it onto [...] the rice bowl.	DAY 2 13:41
	I enter a room where several people are standing around a long table with food. I speak a sentence, but my voice isn't in English...	DAY 2 21:04

Question

DAY 6 15:48

What was the first food I ate along with milk?
A. Banana B. Pancake
C. Eggs D. Cookie
Correct Ans: B

A. Banana
B. Pancake

EgoRAG (EgoGPT)

	[...] I moved the straw slightly with my left hand, placed a finger on the drink, and fell into thought. The table was filled with [...]	DAY 2 13:51
	[...] picked up a spoon, and started stirring the bread in the bowl [...] then picked up a glass of milk from the table [...]	DAY 1 11:26
	[...] Katrina asked, 'Where should I put this?' then said, 'I'll do it.' Tasha reminded, 'There's still a bottle of fresh milk [...]	DAY 1 19:11

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 EgoButler 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

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 highlights key insights into EgoGPT variants for EgoRAG memory bank construction. The use of human caption annotations helps achieve the highest scores, emphasizing the importance of high-quality captions for better retrieval and performance. Among EgoGPT variants, audio-only models perform the weakest, while visual-only models perform better, indicating that audio-only information might not be adequate to solve EgoLifeQA. Combining visual and audio inputs yields the best performance among variants. The consistent improvement with the additional EgoLife Day-1 caption data highlights the importance of incorporating participant-specific information as well as domain-specific generalization.

Qualitative Results Figure 8 showcases the strengths of EgoGPT and the effects of EgoRAG. Compared to Gemini-1.5-Pro, EgoGPT (EgoIT-99K+D1) naturally excels in personalization and generating contextually relevant captions, though with notable limitations. EgoGPT’s speech understanding remains incomplete. It struggles to understand human laughter and emotions, likely due to its reliance on ASR-trained data. Identity recognition is another challenge. Since it was fine-tuned on EgoLife Day-1, it tends to overfit to early observations. For example, if a person wore a blue shirt on Day-1, EgoGPT may misidentify a different person wearing blue later as the same individual, revealing limitations in its current personalization strategy.

EgoRAG complements EgoGPT by retrieving *long-context evidence, but its retrieval mechanism lacks multi-step reasoning. It performs a single-pass search without

iterative refinement or step-by-step reasoning, making it prone to failure when relevant information is missing from direct retrieval. This lack of error tolerance means that if EgoRAG cannot locate supporting evidence, it simply fails to provide an answer rather than reasoning around the missing information.

These findings highlight critical areas for future immediate improvement: enhancing speech comprehension, refining personalization strategies, and incorporating more advanced retrieval reasoning techniques to improve error resilience.

6. Conclusion and Outlook

This work presents EgoLife, a pioneering dataset and benchmark that marks a significant step toward ultra-long egocentric video understanding. Beyond its debut focus, EgoLife unlocks vast untapped potential. The ultra-long collaborative multi-view recordings of six participants provide a unique opportunity to explore synchronized human behaviors, while the ego-exo alignment invites new insights into their natural, everyday dynamics in shared environments. The calibrated cameras, combined with multimodal signals like millimeter-wave radar and WiFi, enriched by detailed annotations, pave the way for diverse approaches to modeling human life across intricate temporal and spatial contexts.

Returning to the vision of building multimodal AI assistants, this work provides explorations we hope will inspire further research in this promising field. *EgoLife is just the beginning—we dream of a future where this work inspires a collective journey, enabling AI to become a truly efficient, empathetic, and transformative companion in human life.*

Acknowledgement

References

- [1] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talatoff, 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](#), [6](#), [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](#), [6](#), [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](#), [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](#), [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. [3](#)
- [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](#), [5](#), [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. [3](#), [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. [3](#), [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](#)
- [15] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Du

- rante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding, 2024. 3
- [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-mmact) database. 2009. 2, 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 egocentric 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. 2
- [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. 2, 6, 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. 2, 6, 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. 2, 6, 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. 2, 6, 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. 2, 6, 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, 6, 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. 25
- [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. 25
- [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, Chunyan 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, 6, 7
- [56] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univ: 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 milm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 3, 6
- [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, Chunyan 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, Xiaoli 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] Weihang 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. Egolpv2: 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, Wenhui 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 Alilegro, 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. 4, 6
- [91] Mahmoud Ashraf. Whisper diarization: Speaker diarization using openai whisper. 2024. 4
- [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. 6, 22, 26
- [94] OpenAI. Gpt-4v(ision) system card, 2023. 6
- [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. 6, 7
- [96] OpenAI. Gpt-4o system card, 2024. 6, 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. 6
- [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. 6
- [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. 6
- [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. 6
- [101] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhi-fang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759, 2024. 6
- [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. 6
- [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. 24
- [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. 24
- [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 Battiatto, Giovanni Signorello, and Giovanni Maria Farinella. Egoch: 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. Egomoc: 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. [25](#)
- [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. [25](#)
- [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. Epic-tent: 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 Shelepor, 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 ego-

- centric 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, Hawkeye Dave, Yang Xuan, Wilson Matthew, Debats Stephanie, Nuengsigkapian Cattalyya, Sharma Astuti, Pandikow Lars, Wang Huisheng, Adam Hartwig, and Sirotenko Mikhail. Sango: 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, Nishiaki 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. [26](#)
- [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 learning multi-agent multi-task activities. 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. Ego-body: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. [26](#)

Contents of Supplementary Material

1. Introduction	2
2. Related Work	2
2.1. Egocentric Datasets & Benchmarks	2
2.2. Long-Context Video Language Models	3
3. The EgoLife Dataset & Benchmark	3
3.1. Data Collection	3
3.2. Data Cleaning	4
3.3. Transcript Annotations	4
3.4. Caption Annotations	4
3.5. EgoLifeQA Annotations	5
4. EgoButler: Agentic Egocentric Life Assistant	5
4.1. System-I: EgoGPT for Clip Understanding	6
4.2. System-II: EgoRAG for Long-Context Q&A	6
4.3. Integration and Synergy in EgoButler	7
5. Experiments	7
6. Conclusion and Outlook	9
A Authorship Statement	17
B Ethical Considerations	18
C Potential Social Impact	18
D EgoLife Dataset Card	19
D.1. Data Capturing	19
D.2. Data Cleaning	19
D.3. Dataset Composition	19
D.4. Dataset Collection Process	19
D.5. Data Preprocessing	20
D.6. Annotations	20
D.7. Dataset Structure	20
D.8. Annotations	20
D.9. Cost Breakdown	20
E Daily Activities	20
F. Details of EgoIT	22
G History of Egocentric Datasets	23
G.1. Egocentric Datasets	23
G.2 Ego-Exo Datasets	26
H Annotation Examples	26

A. Authorship Statement

Jingkang Yang (LMMs-Lab, NTU S-Lab) served as the project lead and director of the entire initiative, overseeing all aspects from the conception of the EgoLife project to its execution. His responsibilities included coordinating the casting and data collection process and organizing and managing all the details such as data cleaning, annotation, model training, evaluation, RAG system construction, paper writing, and public presentation.

Data Collection and Preparation:

- **Shuai Liu (LMMs-Lab, NTU S-Lab)**, **Yuhao Dong (NTU S-Lab)**, **Binzhu Xie (CUHK)**, and **Zitang Zhou (BUPT)** were involved from the project's inception, contributing to the planning and assisting during the EgoLife casting week. **Zitang Zhou** helped in posting and looking for suitable volunteers.
- **Ziyue Wang (NTU S-Lab)** and **Bei Ouyang (IMDEA Networks)** participated in early-stage planning discussions, though they were unable to assist on-site during the casting week.
- **Zhengyu Lin (NTU S-Lab)** provided crucial support in setting up GoPro cameras and calibrating equipment at the EgoLife house. **Zhongang Cai (NTU S-Lab)** and **Lei Yang (NTU S-Lab)** collaborated on developing solutions for first-person and third-person collaborative data collection, contributing both equipment and financial support.
- **Bei Ouyang** and **Joerg Widmer (IMDEA Networks)** contributed to setting up mmWave radars and mmWave signal collection efforts.
- For the English-language subset of EgoLife in Milan, **Jingkang Yang**, **Xiamengwei Zhang**, **Binzhu Xie**, **Bei Ouyang**, **Marco Cominelli (University of Brescia)**, and **Francesco Gringoli (University of Brescia)** all contributed to data collection efforts.
- **Marco Cominelli** and **Francesco Gringoli** were also instrumental in setting up the infrastructure for the WiFi signal data collection for this subset of the project.

Data Cleaning and Annotation:

- **Shuai Liu** took the lead on maintaining and sorting out the raw data. He also organized EgoLife data into the trainable structure using all annotations.
- **Xiamengwei Zhang (CNU)** participated as one of the five external volunteers during the EgoLife casting week, afterward making significant contributions to manage the data annotation team, including all captioning and EgoLifeQA. She also processed and reconstructed the 3D model of the EgoLife house for demo purposes.
- **Hongming Guo (BUPT)** and **Pengyun Wang (ANU)** joined the project after the casting week but made vital contributions to data cleaning efforts.

- **Hongming Guo** worked extensively on multi-view synchronization, desensitization, and other critical tasks, and also played an active role in designing the EgoLifeQA framework.
- **Pengyun Wang** assisted with audio transcript pre-annotation tasks, including diarization, with additional support from **Sicheng Zhang** (**Khalifa University**).
- **Ziyue Wang**, after returning from a leave of absence, made significant contributions to data extraction from VRS files, multi-person VRS synchronization, and exploring multimodal models for multi-view processing.

Model Development, Training, and Evaluation:

- **Yuhao Dong** and **Shuai Liu** led the model training efforts, with substantial support from **Ziyue Wang** and **Zitang Zhou** in organizing and curating the training data.
- **Zitang Zhou** conducted an in-depth review of all relevant egocentric datasets and played a key role in selecting the EgoIT dataset, with valuable assistance from **Binzhu Xie** and **Sicheng Zhang**.
- The development of the EgoRAG framework was carried out by **Hongming Guo**, **Shuai Liu**, and **Sicheng Zhang**.
- **Shuai Liu** and **Hongming Guo** were responsible for defining and implementing the evaluation protocols, including the integration of EgoSchema, EgoPlan, and other elements into the LMMs-Eval framework.

Advising and Discussion:

- **Ziwei Liu** (**NTU S-Lab, LMMs-Lab, corresponding author**) provided regular and decisive guidance throughout the project, offering invaluable resource support that was critical to the successful execution of the project.
- **Bo Li** (**NTU S-Lab, LMMs-Lab**) and **Yuanhan Zhang** (**NTU S-Lab**) contributed extensive expertise and support in model training, providing key insights that greatly enhanced the development and fine-tuning of the model. **Peiyuan Zhang** (**UCSD**) offered valuable insights on long-context video language models, shaping the project's approach to handling complex video data.
- **Fangzhou Hong** (**NTU S-Lab**) provided significant support through his expertise in egocentric research from the perspective of 3D vision, which positioned the dataset for broader impact within the 3D research community.

B. Ethical Considerations

All data collection in this project was conducted in strict compliance with ethical guidelines, ensuring the protection of participants' privacy and the safeguarding of sensitive content. Below, we elaborate on key aspects of our ethical protocols:

- **Permission for Filming Locations:** All filming locations, including private properties such as the villa, were used

with explicit permission from the owners. Written or verbal agreements were established, and prior communications with the owners substantiate this consent.

- **Institutional Review:** The entire data collection process was reviewed and approved by the internal ethics committee of the authors' affiliated institution. While adhering to double-blind review standards, we ensure that all claims align with the necessary ethical documentation and approvals.
- **Handling of Sensitive Content:** Sensitive content was managed with utmost care, employing the following measures:
 - *Blurring of faces and identifiers*: All participant faces were blurred to anonymize identities. Additionally, bystanders' faces and vehicle license plates appearing in the footage were thoroughly blurred.
 - *Audio muting*: Sensitive audio segments containing private or potentially identifiable information were muted to ensure privacy.
 - *Screen privacy*: Frames containing sensitive screen content, such as mobile or computer screens, were reviewed, and any private information was blurred. For example, visible screens displaying passwords or personal data underwent detailed masking processes.
- **Informed Consent:** All participants provided informed consent before the commencement of data collection. They were thoroughly briefed on the purpose, scope, and intended applications of the project, ensuring their voluntary and informed participation.
- **Data Storage and Security:** Raw data was securely stored in accordance with best practices to prevent unauthorized access. Anonymization was applied throughout the dataset to protect participant identities.

By adhering to these rigorous ethical measures, this project ensures the highest standards of privacy, trust, and integrity while advancing AI research.

C. Potential Social Impact

The development of EgoButler and the EgoLifeQA dataset holds significant potential to enhance human-AI interaction, particularly in personalized assistance and context-aware applications. By enabling AI to understand long-term, egocentric perspectives, EgoButler could support daily activities, personal organization, and contextual reminders, improving quality of life, especially for individuals needing consistent support, such as the elderly or those with cognitive challenges.

In educational and professional settings, egocentric AI could facilitate learning, task tracking, and skill development, adapting to individual needs and preferences. However, as this technology integrates more deeply into personal spaces, it is essential to address privacy and ethical considerations to ensure user autonomy and trust. Safeguards for data privacy

and transparency in AI decision-making processes will be key to its positive societal reception.

EgoButler’s advancements may ultimately foster a new era of AI companions capable of supporting individuals in a socially and ethically responsible manner. By promoting real-time, context-aware AI, this work aims to benefit society, encouraging safe, meaningful, and privacy-conscious interactions between humans and AI.

D. EgoLife Dataset Card

The **EgoLife** dataset is a comprehensive collection of ultra-long, multi-participant video recordings captured from both first-person and third-person perspectives, enriched with synchronized multimodal signal data. This ongoing project aims to document human daily activities in natural environments, advancing research in human behavior recognition, multimodal signal analysis, and human-machine interaction.

To date, data has been collected from two distinct environments: one in Beijing, China, and another in Milan, Italy. The Beijing dataset has been fully annotated and synchronized, and fully discussed in the main paper, while the Milan dataset has been collected and will be detailed in the upcoming EgoLife blog series.

D.1. Data Capturing

Curation Rationale The dataset was curated to provide a realistic depiction of human behavior in natural settings, supporting signal-based behavior modeling and exploration of multimodal data synchronization in real-world scenarios. The EgoLife dataset currently has two sessions.

- **Beijing:** Data was collected over seven days, capturing 40+ hours of daily activities. The language of interactions is primarily Chinese.
- **Milan:** A one-day session capturing approximately 6 hours of activity, featuring similar tasks and interactions as Beijing. The language is primarily English, with some Chinese and Italian.

Naming Remarks When we refer to the EgoLife dataset, we refer to the 7-day session in Beijing. We call the one-day EgoLife data from Milan as EgoLife-Milan.

D.2. Data Cleaning

The dataset underwent rigorous data cleaning to ensure quality and remove any sensitive or low-quality segments. All identifiable faces and sensitive license plates were blurred, and audio containing sensitive topics was muted.

D.3. Dataset Composition

Data Instances Each data instance includes:

- First-person video from AI glasses
- Third-person video from fixed indoor cameras

- Synchronized multimodal signal data, including millimeter-wave radars and WiFi signals

Data Fields

- **Video Fields:** Capturing primary visual data from both first- and third-person perspectives.
- **Signal Fields:** Radars and WiFi emitters for spatial and behavior correlation analysis.

Data Statistics

- **Participant Sessions:** Six participants in both datasets; Beijing features 40+ hours over seven days, Milan adds 6 hours in one day.

D.4. Dataset Collection Process

Participants Six volunteers participated in both locations, with varied interactions and activities recorded.

Equipment

- **First-Person AI Glasses:** 6 Aria glasses for continuous video capture from the participant’s perspective.
- **Indoor Third-Person Cameras:** 15 in Beijing, six in Milan (four in living room, two in kitchen).
- **Millimeter-Wave Radars:** Deployed for spatial and movement data collection. Two TI IWR6843 (60GHz) mmWave monostatic radars and corresponding DCA1000 data capture boards in Beijing. Two TI IWR6843 (60GHz) mmWave monostatic radars, one AWR1843 (77GHz) mmWave monostatic radar and corresponding DCA1000 data capture boards in Milan.
- **WiFi Receivers/Emitters:** Deployed for spatial and movement data collection (only in Milan). Three Asus RT-AX82U devices in the living room.

Collection Protocol Participants were asked to perform typical daily activities, with natural interactions captured in various indoor settings.

mmWave Signal Collection and Preprocessing Multiple mmWave radars and corresponding data capture boards are deployed in the corners of rooms. We use monostatic radars, which means both the transmitter and receiver are on the same device. We can estimate the movements and the locations of targets using one single mmWave radar. In this paper, we exploit data capture boards to obtain the raw ADC data streamed from radars. In the post-process of mmWave data, we used the constant false alarm rate (CFAR) detection algorithm to detect dynamic target signals within background noise while distinguishing them from static environmental signals.

WiFi Signal Collection Three Asus RT-AX82U devices are deployed in different corners of the room. One device transmits dummy WiFi frames at an average rate of 20 frames/s; the other two devices filter such dummy frames and collect channel state information (CSI) data independently using the AX-CSI platform. The CSI, measured by each receiver for each incoming WiFi frame, estimates the WiFi channel frequency response between the transmitter and the receiver. Specifically, we transmitted over the WiFi channel regular 802.11ax frames with 160 MHz bandwidth and 4x4 multiple-input multiple-output (MIMO) configuration. Hence, the CSI extracted by each receiver from every frame consists of 2048 orthogonal subcarriers and 16 separate spatial streams, i.e., a total of 2048×16 complex (real and imaginary parts) data points per frame.

D.5. Data Preprocessing

Multimodal Signal Extraction Signal data, including radar and WiFi, were extracted and aligned with video data to create a comprehensive multimodal dataset.

Multi-view Synchronization Video and signal data from multiple sources were synchronized using timestamps for cohesive analysis.

De-identification Process All faces and sensitive visual data were blurred. Any sensitive topics in audio were muted to protect participant privacy.

Audio Processing Audio was processed to mute sensitive information and enhance clarity for Q&A annotations.

D.6. Annotations

- Annotation Process:** Initially generated with GPT for Q&A, followed by human refinement for relevance. Activities and events are annotated across two levels: fine-grained and integrated.
- Annotation Types:** Includes event/activity labels and Q&A annotations to support contextual and semantic analysis of recorded scenes.

D.7. Dataset Structure

Data Splits Data is divided by location:

- **Beijing Dataset:** Multi-day dataset in Chinese.
- **Milan Dataset:** Single-day dataset, primarily in English.

File Formats Data files are stored in standard formats for easy accessibility:

- Video+Audio: MP 4
- IMU: CSV
- Gaze: CSV
- Radar Signal Data: CSV

- WiFi Signal Data: HDF 5
- Annotations: JSON

D.8. Annotations

- Annotation Process:** Initially generated with GPT for Q&A, followed by human refinement for relevance. Activities and events are annotated across two levels: fine-grained and integrated.
- Annotation Types:** Includes event/activity labels and Q&A annotations to support contextual and semantic analysis of recorded scenes.

D.9. Cost Breakdown

As the first step toward a realistic egocentric life assistant, we intentionally started with a narrow setting to build a strong foundation, sacrificing some generalizability (e.g., single language/scenario). However, we see great value in expanding the project while encouraging community contributions. To support scalability, we report the data collection cost breakdown as below. Finding a reliable annotation team took two months and five trials, and this partnership will continue for future EgoLife versions.

Before Data Collection		During Data Collection		After Data Collection			
Items	Q.	USD	Items	USD	Items		
Hard Drives (4T)	10	420	Housing Rent Expenses	2,250	Caption Annotation	60	3,000
Laptop Rental (10 days)	10	200	Volunteer Allowance	1,380	Speech Transcript	50	2,800
GoPro Rental (10 days)	15	200	Equipment Expenses	1300	QA Annotation	2m	2,760
SD Card	15	320	Meal Expenses	690	Synchronization	3m	1,400
Power Bank	20	280	Transportation	300	Desensitization	1m	1,400
Accessories	-	280	Others	150	Translation	10	700
Sum	-	1,700	Sum	6,070	Sum	-	12,060

Q. means quantity, Dur. means duration, i.e., the number of days needed for one annotator to complete the task.

E. Daily Activities

Day 1: Planning and Initial Preparations On the first day of our week-long experiment, the six participants began by holding a planning meeting to discuss the primary goal of organizing a World Earth Day-themed party on the sixth day. This meeting set the stage for the following days, as we outlined the key tasks and responsibilities for everyone.

In the afternoon, we embarked on the first round of grocery shopping. This was essential not only for ensuring we had enough supplies to sustain ourselves throughout the week but also to gather ingredients for the meals we planned to prepare during the experiment.

The evening was spent showcasing our culinary skills. Each participant took charge of preparing dishes using the fresh ingredients purchased earlier in the day. This collaborative cooking session helped foster camaraderie among the group and provided an enjoyable conclusion to the first day of activities.

Day 2: Dance Practice and Room Decorations The second day was dedicated to creative and physical activities, laying the groundwork for the Earth Day party. In the morning, we brainstormed ideas for a group dance performance

to showcase during the party. This involved watching online videos, selecting suitable choreography, and assigning roles. At the same time, some participants started crafting handmade decorations to align with the Earth Day theme. These decorations were intended for both personal rooms and the shared party space.

In the afternoon, we moved from planning to action, practicing the dance routine based on the morning's decisions. The rehearsals were filled with energy and laughter, as everyone contributed to refining the choreography. Meanwhile, others focused on enhancing the visual appeal of the house by decorating rooms with eco-friendly and Earth-themed designs.

After the creative and physical exertions of the day, we enjoyed a hotpot dinner together in the evening. This communal meal was followed by informal discussions, during which participants took turns explaining their decoration ideas for their respective rooms and how these designs aligned with the Earth Day theme. This exchange of ideas not only inspired creativity but also reinforced the shared vision for the event.

Day 3: Games, Outdoor Exploration, and a Feast The third day began with a fun and lighthearted game involving taste-testing various brands of water. Each participant attempted to identify the brand of water based solely on taste. This game not only served as an engaging activity but also established a points system that would later determine the order of gift exchanges during the party.

In the afternoon, we ventured outdoors for some fresh air and inspiration. Initially, we planned to film a vlog during this outing, but the focus shifted to simply enjoying nature and gathering ideas. We strolled through a nearby park, soaking in the scenery, and later stumbled upon an arcade where we indulged in games like claw machines.

The evening turned into a culinary extravaganza. After another round of shopping for fresh ingredients, we prepared a grand meal together, featuring a variety of dishes. The feast included barbecue, homemade desserts like cakes, and other delightful creations. The shared cooking and dining experience brought everyone closer and added to the festive atmosphere of the day.

Day 4: Seasonal Festivities, Decorations, and a Mishap

The fourth day began with a special nod to the calendar. As it coincided with a significant seasonal event, we marked the occasion by ordering and enjoying a traditional breakfast associated with the day. After breakfast, participants focused on tidying up the house, cleaning up after the previous day's activities, and continuing their personal room decorations for the Earth Day theme. The arrival of packages containing decorative items added momentum to the effort.

In the afternoon, some participants ventured out to a nearby café that allowed interaction with animals, particularly dogs. While this was meant to be a relaxing activity, one participant was bitten by a dog, necessitating a trip to get vaccinated in the evening.

Meanwhile, others remained at home to further enhance their room decorations and refine plans for the party. Evening activities included a mix of lighthearted entertainment, such as singing to lift spirits, and creative tasks like making desserts. To wrap up the day, everyone gathered to finalize the details and schedule for the Earth Day party, ensuring the plan was clear and cohesive.

Day 5: Final Preparations The fifth day was all about wrapping up the remaining tasks before the big Earth Day party. The morning was a flurry of activity as participants worked on unfinished decorations and handmade crafts, ensuring everything was aligned with the party's theme. While eating and staying energized remained essential, the main focus was on completing creative tasks.

In the afternoon, we went on the final grocery run to ensure we had enough supplies to host our guests the next day. Later in the evening, we picked up packages containing key decorative items and materials that had arrived just in time. The night was dedicated to fine-tuning the room setup and conducting one last round of discussions about the party's schedule and activities.

Day 6: The Earth Day Party The sixth day marked the culmination of all our efforts: the Earth Day party. The morning was a race against the clock as we completed final cleaning and decoration touches. In the afternoon, we welcomed our guests, guiding them to the venue.

The party started with an opening segment, followed by a screening of a short video montage we had prepared earlier in the week. Next was a Q&A session where participants and guests could earn "EgoCoins," a virtual currency we had created for the event. These coins could be used during a lively auction featuring handmade crafts and small items contributed by the organizers and guests alike.

After the auction, guests were given a guided tour of each participant's themed room, showcasing the hard work and creativity that had gone into decorating them.

The evening was a celebration of connection and joy. We enjoyed a barbecue, sang songs, and engaged in casual conversations, creating a relaxed and vibrant atmosphere to cap off the day.

Day 7: Cleanup and Farewell The final day was dedicated to dismantling the decorations and cleaning up the house. Since the house was a rental, we made sure to restore it to its original condition. Participants carefully packed

away personal belongings and bid farewell to the themed rooms they had worked so hard to create.

In the evening, we shared a final meal together, reflecting on the experiences of the past week and saying our goodbyes. With heartfelt farewells, we closed this unique chapter of our journey, leaving with unforgettable memories of a week spent living, creating, and celebrating together.

F. Details of EgoIT

To construct the instruct tuning data, EgoIT, we carefully curated a diverse set of egocentric datasets, strategically chosen to ensure comprehensive coverage across a spectrum of activities, environments, and interactions. This diversity is crucial for training robust and generalizable egocentric models. Ego4D [5] provides extensive daily-life activity videos across multiple scenarios, offering a broad foundation for egocentric AI research. HoloAssist [29] focuses on human-object interactions in augmented reality settings, contributing insights into AR-based tasks and interactions. EGTEA Gaze+ [26] emphasizes gaze tracking and action recognition, aiding in understanding attention and intention during activities, crucial for anticipating user needs and providing proactive assistance. IndustReal [28] targets industrial and professional tasks, addressing the specific needs of professional environments by adding specificity to workplace scenarios. EgoTaskQA [93] is designed for egocentric question answering, enhancing model's task-based reasoning capabilities, crucial for understanding instructions and providing relevant responses. EgoProceL [27] focuses on procedural learning and task segmentation, allowing the model to learn step-by-step guidance and understand the temporal structure of complex activities. Charades-Ego [25] employs a randomized action selection methodology to collect a diverse and highly life-relevant dataset on a global scale, improving the model's ability to generalize across various cultural contexts. Epic-Kitchen [4] offers detailed annotations of cooking-related activities, strengthening comprehension of intricate, multi-step tasks in domestic environments. Finally, ADL [24] provides insights into routine human behaviors and object interactions, ensuring models are equipped for assisting in everyday tasks. By integrating these datasets, EgoIT aims to create a balanced and comprehensive training resource, enabling the development of more robust and versatile egocentric AI applications. The prompt to generate Q&A data is shown as follows.

System Message:

QA pairs prompt:

You are a question-answer generation
 ↳ assistant. You should help me
 ↳ generate some QA pairs with the
 ↳ reference of the "text" caption

↳ I provide you. There are also
 ↳ some instructions that you might
 ↳ follow:

1. Your question for the Q-A pairs
 ↳ should be multi-dimentional, for
 ↳ example you can brainstorm
 ↳ question from aspects like
 ↳ reasoning, planning, activity
 ↳ localization etc.
2. Your Q-A pairs should be easy to
 ↳ respond, even by a human, which
 ↳ means you should focus more on
 ↳ the fact of the caption rather
 ↳ than the subjective feeling or
 ↳ aspects.
3. Your question should be general
 ↳ enough, and the length of both
 ↳ question and answer can be
 ↳ various.
4. Make sure that the QA pairs you
 ↳ generated can be confidently
 ↳ answered.
5. For each Index, kindly give me more
 ↳ than 7 QAs.
6. Try to generate some answers simply
 ↳ with "No" or "Yes".
7. Generate some answers which are
 ↳ "No", the question for "No"
 ↳ answer can be made up.
8. Generated QA should be visually
 ↳ conducted rather than hear or
 ↳ sense. (E.g. You can't see you
 ↳ are laughing, try to use visible
 ↳ predicates)
9. The format of your respond should
 ↳ be:

Index x

Timestamp: xxx - xxx

Q: xxx

A: xxx

Q: xxx

A: xxx

Q: xxx

A: xxx

...

Here are some types of answer you may
 ↳ generate for your reference:

1. Descriimative question (Yes or No
 ↳ questions or choice):
- Q: In this video, am I playing board
 ↳ games with other people?
- A: yes

Q: Am I using a machine in the video?
 A: no
 Q: What is this place in the video,
 → forest or sea?
 A: Forest.
 Q: Where am I, indoor or outdoor?
 A: Outdoor.
 Q: Is the thing holding in my right
 → hand made of plastic or not?
 A: It is not made of plastic
 Q: What gender am I most likely to be?
 A: Women.

2. Descriptive questions:

Q: What are the main ingredients and
 → tools used during the video, and
 → how do they contribute to the
 → goal of the activity?
 A: The main ingredients used in the
 → video are peas, water, and salt.
 → the main tools used are a
 → measuring cup, a pan, and a
 → spoon."
 Q: What am I doing?
 A: Ironing clothes.
 Q: What am I holding in my right hand?
 A: A brush.
 Q: How do I break the item I'm holding
 → in my left hand and pour it into
 → the bowl?
 A: Tap it firmly against the edge of
 → the bowl to crack the shell and
 → then use your fingers to gently
 → pull the two halves apart over
 → the bowl.

3. Make predictions base on current → and future timestamps:

Q: will watermelon be visible to the
 → other person after the person's
 → next action?
 A: yes
 Q: What will I do next?
 A: Open the car door.
 Q: What will I put in the washing
 → machine?
 A: Clothes.
 Q: What will the status of fork change
 → to if the actor do the first
 → action in the video in the
 → future?

A: on top of plate
 Q: What will I do?
 A: Take out the mushrooms.

4. Reason task:

Q: What is the use of the object in my
 → left hand?
 A: Serving food
 Q: What's the use of the object in my
 → right hand?
 A: Eating food

Now, I will give you some
 → informations! You should mimic
 → the tune of sample QAs and help
 → generate some general questions
 → following the required format to
 → finish the QA pairs.

G. History of Egocentric Datasets

G.1. Egocentric Datasets

Following A1, early egocentric datasets were mainly small in scale, focusing on specific human activities and targeting recognition tasks. EgoActions [104] is a sports-focused egocentric dataset with 8 videos, annotated with activity labels. VNIST [105] captures ego-motion during walking to work, with 31 videos annotated with location and novelty labels for novelty detection. ADL [106] consists 10 hours of video annotated with activity labels, bounding-box tracks of all visible objects, and interaction annotations for action and object recognition. Social Interactions [17] is a dataset of 42 hours of video annotated with interaction types for detecting and analyzing social interactions. UT-Ego [18] is one of the earliest egocentric dataset that incorporates gaze modality and text annotations, with a collection of 37 hours of first-person videos annotated with video summarization and object segmentations. JPL-Interaction [19] features 57 videos of human interactions for action recognition tasks. BEOID [107] focuses on task-relevant objects and their modes of interaction from multi-user egocentric video annotated with gaze and action labels. HUJI EgoSeg [108] contains 65 hours of videos annotated with activity labels and timestamps. FPPA [109] includes 591 videos of same daily-life activities performed by different subjects. Stanford ECM [110] contains 31 hours of videos annotated with activity classes and metabolic equivalents of task for activity recognition and energy expenditure estimation. OST [111] features 57 sequences of egocentric videos annotated with object labels and gaze points for object search tasks using eye-tracking data. The THU-READ dataset [112] is composed of 1920 RGB-D sequences captured by 8 participants who performed 40 different daily-life actions. DoMSEV [113] is



Figure A1. The Overview of Egocentric Datasets. The figure summarizes the domain, modality, annotation type, release time, dataset statistics, and other aspects of datasets, providing a comprehensive view of existing egocentric datasets.

an 80-hour egocentric dataset designed for fast-forwarding videos while retaining relevant information, with annotations for scene and activity labels. IU ShareView [114] provides 9 paired first-person videos (5–10 minutes each) annotated with bounding boxes and person IDs for person segmentation and identification. EgoCart [115] captures shopping activities in retail stores, with camera pose ground truths and class labels for indoor localization and shopping cart detection. EGTEA Gaze+ [26] presents egocentric cooking activities recorded with detailed gaze tracking. DR(eye)VE [22] contains videos with eye-tracking annotations for predicting the driver’s focus of attention during driving tasks. More egocentric datasets have expanded beyond specific activity recognition tasks to explore a broader range of topics, reflecting the diverse and multidisciplinary nature of egocentric vision research. EgoVQA [116] is a question-answering dataset with 600 QA pairs and 5,000 frames aimed at VideoQA tasks using egocentric video. Ego-CH [117] focuses on cultural heritage videos annotated with environment labels and object retrieval labels for localization in cultural sites. EgoCom [118] contains 38.5 hours annotated with speaker labels and word-level transcriptions for understanding human communication and turn-taking. You2Me [119] is a dataset for 3D body pose estimation from egocentric video, featuring skeleton poses and activity labels. Ego-

Deliver [120] contains 5,360 videos from takeaway riders annotated with action, goods, and event labels for activity detection and recognition. Touch and Go [121] combines tactile sensor data with egocentric videos for visuo-tactile feature learning and material recognition in natural environments. HOI4D [122] is a 4D dataset with 2.4M frames of indoor human-object interactions annotated for action segmentation, 3D hand pose, and object tracking. EgoObjects [123] is a large-scale egocentric dataset with 9K videos annotated for instance-level and category-level object detection, aiming to enhance continual learning. Arial Digital Twin [124] focuses on AR/VR applications involving digitized environments and egocentric interactions. WEAR [125] is a sports-related dataset with 15 hours of videos annotated with activity labels for activity recognition tasks. EGOFALLS [126] is a dataset for fall detection, featuring 10,948 video samples annotated with activity and environment labels. While earlier datasets had limitations in certain aspects, more recent ones have made progress in terms of scale and generality. The EPIC-KITCHENS dataset [4] was a pioneer in large-scale egocentric action recognition, focusing on kitchen environments. Ego4D [5] expanded beyond this, covering a wider range of daily activities and becoming one of the most widely-used egocentric datasets due to its massive scale. Several datasets have since built upon EPIC-

KITCHENS and Ego4D. For instance, TREK-150 [127] selected 150 videos from EPIC-KITCHENS and added bounding boxes for object tracking, while VISOR [31] incorporated 36 hours of EPIC-KITCHENS footage and provided dense hand masks and object labels. N-EPIC-KITCHENS [32] enhanced all EPIC-KITCHENS videos by adding event annotations. EpicSoundObject [128] filtered out silent videos from EPIC-KITCHENS, resulting in 13,000 frames with bounding boxes of sounding objects. VOST [33] used 4 hours of video from EPIC-KITCHENS and Ego4D, focusing on complex object transformations and providing dense instance masks. EgoClip [30] filtered 2,900 hours of video from Ego4D that lacked narrations, adding timestamp-level narrations. EgoSchema [11] took long-form videos from Ego4D and created multiple-choice question-answer pairs, making it a popular resource for long video understanding. PVSG [129], consisting of 111 videos from Ego4D and EPIC-KITCHENS, appended frame-wise panoptic segmentation masks.

There is a specific set of datasets focusing on procedural learning in assembly or instructional scenarios, emphasizing the identification of key steps. EPIC-Tent [130] offers 5.4 hours of tent assembly videos along with action labels. MEC-CANO [131] includes 20 videos where participants build a motorbike model. Assembly101 [132] simulates an industrial environment, comprising 513 hours of assembly and disassembly videos of toy vehicles, captured from multiple perspectives. AssistQ [133] features 100 videos and 529 QA pairs designed for AI assistants to learn from instructional videos and provide step-by-step guidance from the user's perspective. EgoProceL [27] centers on procedural learning, providing 62 hours of video where people perform 16 tasks, annotated with step labels and timestamps. ENIGMA-51 [134] consists of 22 hours of video in an industrial setting, where 19 participants followed instructions to repair electrical boards. HoloAssist [29] introduces human interaction by detecting collaboration during manipulation tasks. Lastly, InsudtReal [28] includes 84 toy assembly videos, focusing on recognizing the correct sequence and completion of procedural steps. EgoYC2 [135] is an egocentric instructional video dataset, re-recording YouCook2 [136] cooking videos with procedural captions for video captioning tasks.

Some egocentric datasets focus specifically on hands and their interactions with objects, advancing the understanding of hand-object interactions, gesture recognition, and hand pose estimation. Handled Objects [137] features 10 videos of daily object manipulation activities, annotated with object labels, hand segmentations, and object-ground segmentations for egocentric object recognition. EDSH [138] provides egocentric videos with pixel-level hand masks, designed for detecting hands under challenging conditions such as rapid illumination changes. EgoHands [20] is a dataset of 130,000 frames (4,800 with pixel-level hand masks) for egocentric

hand detection in tabletop games. EgoGesture [139] provides large 24,000 gesture samples (3M frames) annotated with gesture class labels and temporal indices for gesture detection. EgoDexter [140] contains 3,190 frames of hand-object interactions with depth and fingertip position annotations for hand pose estimation. FPHA [141] consists of 1175 videos with action categories and hand-pose annotations for hand pose estimation and action recognition. H2O [142] is a large dataset of synchronized RGB-D frames annotated with hand and object poses for hand-object pose estimation. EgoPAT3D [143] is a household activity dataset featuring 10-hour videos, annotated for 3D action target prediction in human-robot interaction contexts. EgoHOS [144] provides a hand-object segmentation dataset annotated with interaction labels, integrating data from Ego4D [5], EPIC-KITCHENS [4], and THU-READ [112]. AssemblyHands [145] is a 3D hand pose estimation dataset sampled from Assembly101, featuring 3.0M annotated images for hand-object interaction tasks.

Recently, more egocentric-related research has emerged, further enriching the field with diverse datasets, benchmarks, and methodologies. EgoVid-5M [146] introduces a large-scale dataset of 5 million egocentric video clips, facilitating advancements in video generation. In hand-object interaction studies, HOT3D [147] focuses on 3D tracking from multi-view egocentric videos, while EgoPressure [148] provides hand pressure and pose estimation data. Activity recognition and feedback have also progressed, with ExpertAF [149] generating expert feedback from videos, and EgoSurgery-Tool [150] and EgoSurgery-Phase [151] contributing surgical tool detection and phase recognition datasets.

Benchmarks such as EgoPlan-Bench2 [152] for multimodal large language model planning and MomentSeeker [153] for moment retrieval in long videos enhance evaluation frameworks. Vision-language integration is also expanding, with SPHERE [154] identifying spatial blind spots in models and EgoTextVQA [155] advancing egocentric scene-text-aware video question answering. Research into spatial cognition and navigation has been supported by SANPO [156] for human navigation datasets, studies exploring out-of-sight memory in egocentric perception [157], and MLVU [158], which benchmarks multi-task long video understanding.

Quality assessment and tracking improvements are reflected in ESVQA [159]'s perceptual evaluation of spatial videos and EgoPoints [160]' advances in point tracking. Personal assistance systems benefit from EgoMe [161]'s "follow me" capabilities in real-world settings and BioVL-QR [162]'s biochemical vision dataset using micro QR codes. Additionally, detecting activities of daily living in egocentric videos has been explored in [163], focusing on hand use in outpatient neurorehabilitation settings. Lastly, mistake detection and predictive modeling have been explored in EgoOops [164],

which detects procedural errors in egocentric videos, and “Acquisition through My Eyes and Steps” [165], which develops a predictive agent model for egocentric environments. We acknowledge these important contributions, which have significantly shaped the landscape of egocentric video research and continue to inspire developments such as Ego-Life.

G.2. Ego-Exo Datasets

Early efforts like PEV [166], CMU-MMAC [16] and CharadesEgo [25] started to focus on capturing both egocentric and exocentric video. PEV provide paired video of interacting people in both first and third view, annotated with action labels for action recognition in human interactions. CMU-MMAC records participants cooking five different recipes in a lab kitchen using multiview setups, while CharadesEgo focuses on home activities annotated with free-text descriptions. In CharadesEgo, videos are captured sequentially from egocentric and exocentric perspectives, resulting in unsynchronized footage with non-exact activity matches. LEMMA [167] expands on this by featuring multi-agent, multi-task activities in 14 kitchens and living rooms. EgoTaskQA [93] then build a video QA dataset based on LEMMA, annotated with object states and relationships for descriptive, predictive, and counterfactual reasoning tasks. Homage [168] contributes 30 hours of egocentric and exocentric video, documenting 27 participants engaged in household tasks such as laundry. Multi-Ego [169] offers 12 hours of multi-view video and includes selected shots that best represent each video, specifically for video summarization tasks. EgoBody [170] captures human motions during social interactions from both third-person and egocentric perspectives, aiming to estimate human pose, shape, and motion.

While most ego-exo datasets focus on specific scenarios, the following datasets offer larger-scale data spanning a wider range of domains. EgoExoLearn [9] offers 120 hours of egocentric videos simulating the process of learning from human demonstrations through exocentric demonstration videos. Ego-Exo4D [8] simultaneously captures egocentric and exocentric perspectives of skilled human activities, producing long-form recordings with totaling 1,286 hours of video.

H. Annotation Examples

To facilitate the review and verification of annotations, all caption annotations are stored in the SRT format. This format is widely compatible with video software, allowing annotations to be overlaid on videos for direct alignment and validation by human reviewers. The ease of integration with video playback ensures that annotations can be efficiently reviewed and adjusted for accuracy.

Each SRT file is composed of the following components:

- **Interactive instance:** This section captures the objects present in the scene during the specified time interval. It provides a detailed account of the key objects interacting with or being relevant to the protagonist.

- **Action:** This part records the actions or interactions of the protagonist with the identified objects during the corresponding time period. It provides granular details about the behaviors and activities observed.

- **Merged Caption:** This annotation consolidates information from multiple modalities, integrating text, visual data, and audio content. The *Merged Caption* is a comprehensive description that combines:

- The output of **Visual Captioning**, which summarizes the scene based on visual elements captured in the video.
- The output of **Audio Captioning**, which incorporates spoken dialogue or relevant sound events.
- Additional contextual details to provide a coherent, multi-modal narrative of the scene.

The *Merged Caption* thus represents a holistic understanding of the scene, leveraging both visual and auditory cues.

Each entry in the SRT file corresponds to a specific time interval in the video. One concrete example is like below.

```

1
00:00:00,466 --> 00:00:08,800
Action: Holding, walking past, looking
Interactive instance: Phone,
    ↪ staircase, Jack
Merged caption: I was holding a phone
    ↪ and saw Jack walk past me and go
    ↪ up the stairs.
Visual-audio caption: I was holding a
    ↪ phone in my right hand, standing
    ↪ at the living room entrance, and
    ↪ saw Jack walk past me and go up
    ↪ the stairs. I heard Alice say,
    ↪ ``Shouldn't you invite me?'' and
    ↪ I responded, "Where is it
    ↪ charging?"
```



```

2
00:00:08,800 --> 00:00:12,066
Action: Turning left, turning right,
    ↪ walking
Interactive instance: None, none,
    ↪ living room
Merged caption: I turned left, then
    ↪ right, and walked toward the
    ↪ living room, where I saw several
    ↪ people sitting around a table.
Visual-audio caption: I turned left,
    ↪ then right, and walked toward
    ↪ the living room. Several people
```

↪ were busy around the table in
↪ the living room, seemingly
↪ preparing something. The table
↪ was covered with various items,
↪ including cardboard boxes and
↪ small scattered objects. Someone
↪ in green clothes was organizing
↪ things, while others sat at the
↪ table, watching her intently.

3

00:00:12,266 --> 00:00:16,933

Action: Walking, picking up, looking
Interactive instance: Dining table,
↪ power bank, power bank

Merged caption: I walked left past the
↪ dining table, picked up a power
↪ bank, and checked its battery
↪ level.

Visual-audio caption: I walked left
↪ past the dining table, picked up
↪ a power bank from the table, and
↪ checked its battery level. The
↪ dining table was covered with
↪ various items, including tape,
↪ scissors, and some unopened
↪ packages. Nearby, several people
↪ were busy preparing things: one
↪ person was checking their phone,
↪ while another was organizing
↪ items on the table.

4

00:00:17,866 --> 00:00:21,666

Action: Walking to, turning around,
↪ walking out, heading to

Interactive instance: My room, none,
↪ room, Shure's room

Merged caption: I walked to my room,
↪ turned around, walked out, and
↪ headed to Shure's room.

Visual-audio caption: I walked into my
↪ room, which was filled with
↪ electronic equipment and several
↪ monitors. I turned around and
↪ left the room, heading to
↪ Shure's room. Inside, there was
↪ a messy bed and desk covered
↪ with various documents and a
↪ laptop.

examples and qualitative results.