Contents of Supplementary Material

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 (Politecnico di Milano, Italy)**, and **Francesco Gringoli (University of Brescia, Italy)** 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. Potenial 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 Prepocessing 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: `MP4`
- IMU: `CSV`
- Gaze: `CSV`
- Radar Signal Data: `CSV`
- WiFi Signal Data: `HDF5`
- 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

items	Q.	$USD
Hard Drives (4T)	10	420
Laptop Rental (10 days)	10	200
GoPro Rental (10 days)	15	200
SD Card	15	320
Power Bank	20	280
Accessories	-	280
Sum	-	1,700

During Data Collection

items	$USD
Housing Rent Expenses	2,250
Volunteer Allowance	1,380
Equipment Expenses	1300
Meal Expenses	690
Transportation	300
Others	150
Sum	6,070

After Data Collection

items	Dur.	$USD
Caption Annotation	60	3,000
Speech Transcript	50	2,800
QA Annotation	2m	2,760
Synchronization	3m	1,400
Desensitization	1m	1,400
Translation	10	700
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. Descrimiative 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. Discriptive 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 earlist 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] focus 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 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

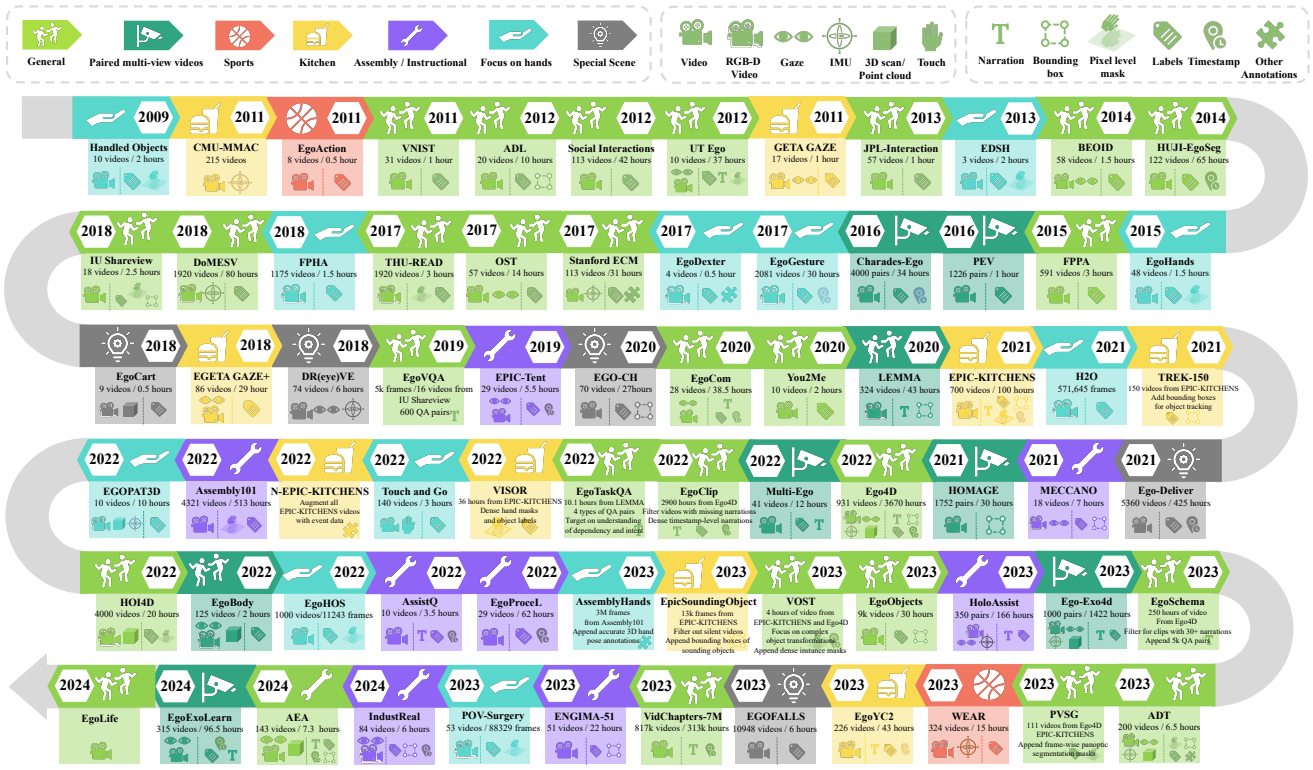

Figure A1. **The Overview of Egocentric Datasets.** The figure summarizes the domain, modality, annotation type, release time, dataset statistics, and other aspects of datasets, providing a comprehensive view of existing egocentric datasets.

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] focus 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. EpicSoundingObject [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. MECCANO [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 EgoLife.

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.
```

Please visit the EgoLife webpage (https://egolife-ai.github.io/) for additional annotation examples and qualitative results.