# EgoEverything: A Benchmark for Human Behavior–Inspired Long-Context Egocentric Video Understanding in AR Environment

Qiance Tang[1†], Ziqi Wang[1†], Jieyu Lin[2], Ziyun Li[2], Barbara De Salvo[2], and Sai Qian Zhang[1‡]

[1]New York University [2]Meta Reality Labs
† Equal contribution. ‡ Corresponding author.

**Abstract.** Long-context egocentric video understanding has recently attracted significant research attention, with augmented reality (AR) highlighted as one of its most important application domains. Nevertheless, the task remains highly challenging due to the need for reasoning over extended temporal contexts and diverse, unstructured activities. Although several benchmarks exist, most egocentric datasets rely on human-worn cameras and focus mainly on visual content, with limited consideration of underlying user behavior when forming video-related queries. EgoEverything is a benchmark that explicitly considers human behavior by leveraging human attention signals, abstracted from gaze data, when generating questions. It comprises over 5,000 multiple-choice question–answer pairs, spanning more than 100 hours of video. By integrating human attention signals during question generation, it more faithfully captures natural human behavior and offers a realistic evaluation setting for long-context egocentric video understanding in AR.

**Keywords:** Long-context Egocentric Video Understanding · Augmented Reality · Human Attention Modeling

## 1 Introduction

Augmented Reality (AR) is emerging not only as a novel user interface but also as a machine learning (ML) platform that integrates embodied experiences such as sensing, perception, memory, and speech. By aligning digital and physical realms, AR enables real-time information extraction and contextual decision-making, transforming domains including healthcare [13, 21], education [4, 48], and industry [14, 29].

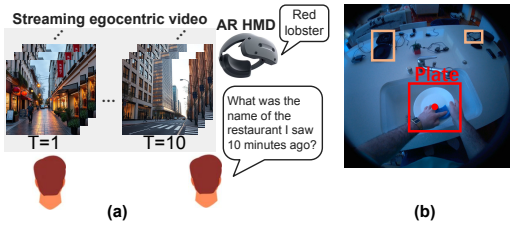


**Fig. 1:** (a) Real-life AR LEU scenario. (b) User attention variation.

Beyond immersive applications, AR devices generate continuous multimodal data streams that are invaluable for ML research. Equipped with cameras, eye-

tracking, hand-tracking, and motion sensors, they capture high-bandwidth visual, auditory, and behavioral signals in real time. These heterogeneous streams pose significant challenges but also offer new opportunities for long-context ML models to capture correlations between signals, user attention, and the surrounding environment across extended timescales, enabling more effective assistance for AR users in everyday scenarios.

As illustrated in Figure 1 (a), consider an AR head-mounted display (HMD) used for driving navigation. The device continuously encodes multimodal inputs through ML models. When the user briefly glances at a restaurant, this gaze event must be accurately linked to the corresponding visual features. Later, answering a query such as "What was the name of the restaurant we passed 10 minutes ago?" requires an episodic memory module that can index temporal embeddings, retrieve the relevant instance, and integrate it with a machine learning model (e.g., Vision-Language Model (VLM)) for analysis. This paradigm transforms AR from a passive data collector into an intelligent personal assistant powered by long-context ML. In everyday life, AR could enable superhuman memory, helping users retrieve details such as where they left their keys. In education, students could revisit past demonstrations or experiments to reinforce learning. In healthcare, surgeons might query the system to recall the exact moment when a critical anatomical landmark appeared during a procedure. Building toward this vision, long-context egocentric video understanding (LEU) has become an increasingly active area of research in the ML community. Long egocentric recordings capture extended daily activities and interactions, requiring models to reason over temporal dependencies and multimodal signals spanning minutes or hours. To evaluate progress, a growing set of benchmarks has been introduced [11, 39, 49, 59], each designed to test how well models can recall, integrate, and reason over such extended sequences. While these benchmarks represent important advances, they largely emphasize generic video-based queries and fall short of capturing the **human-centric** and **attention-guided** nature of real AR usage, where questions are often grounded in what the user was attending to at a given moment. These limitations are summarized as follows:

**Questions Not Reflecting Human Attention:** Existing benchmarks rarely consider user attention when designing queries, creating a mismatch with real-world usage. In practice, people tend to ask about objects or events they have looked at, fully or partially, where their attention was directed. Current datasets instead emphasize generic questions about visual details or scene overviews, and fail to reflect human inquiry patterns.

**Questions Not Framed in Natural Language:** Prior benchmarks often rely on rigid, template-based question generation that does not align with authentic human questioning. For example, prompts such as "Is the light off in the video?" frequently appear, but they are uncommon in daily use. In contrast, real users are more likely to ask context-specific, attention-driven questions such as "Did I forget to turn off the lights?"

**Questions Not Aligned with The Moment of Interaction:** Most benchmarks restrict questioning to occur before or after a clip has ended. However,

users typically pose questions **during** ongoing interactions, requiring real-time reasoning over partially observed streams.

To address these limitations, we present *EgoEverything*, a benchmark for LEU that simulates real-life interactions with AR glasses. Collecting egocentric video in realistic AR scenarios and manually authoring multiple choice questions is both labor and time intensive. Annotators must repeatedly review videos, verify fine grained details, craft challenging queries, and refine phrasing to approximate natural user language. As a result, many prior works resort to template based question generation [31,39,50], which lowers cost and improves consistency but fails to capture how AR users actually ask questions. In contrast, EgoEverything is constructed through a novel Visual Question Answering (VQA) generation pipeline that leverages multiple AI agents to produce questions aligned with authentic human questioning patterns. We further introduce an attention inspired sampling strategy that selects question targets based on real gaze, enabling the benchmark to include both attention driven queries and detail oriented ones outside the user's focus. This design raises task difficulty while more closely matching real world AR query behavior. Finally, we incorporate **comprehensive human review** to enhance question quality and reliability. Specifically, our contribution can be summarized as follows:

- To incorporate human attention into question generation for LEU benchmark, EgoEverything involves an attention-inspired sampling strategy that selects targets based on real gaze, enabling both attention-driven and detail-oriented queries.
- We propose a VQA pipeline with multi-agent question generation and attention inspired sampling based on real gaze, producing both attention driven and detail oriented queries. Rule based filtering and human review further ensure quality and reliability.
- EgoEverything comprises over 5,000 multiple-choice question–answer pairs across more than 100 hours of video. Evaluation on several cutting-edge VLMs reveals consistently lower performance on EgoEverything, highlighting current limitations of VLMs in handling real-life AR LEU scenarios.

## 2 Background and Related Work

**Augmented Reality** (AR) devices are rapidly maturing into always-on, wearable interfaces that bridge virtual content with the physical world through lightweight headsets (e.g., Meta Aria [18]). Modern units typically include front high-resolution camera ($1408 \times 1408$ on Meta Aria glasses) and support natural interaction via hand gestures and human gaze. Crucially, today's headsets can track gaze reliably in real time [25], providing a solid engineering basis for attention modeling and gaze-conditioned interaction.

Unlike phones or desktops, AR headsets are designed and expected to be worn for extended periods and operate under tight battery, thermal, and memory budgets [18, 22]. On these devices, compute is embedded (CPU/GPU), but

RAM is limited and often shared between CPU and GPU (e.g., on Meta Quest $3 \approx 8$ GB [27]), constraining model size and throughput. Meanwhile, a practical AR assistant must continuously process long video streams while the user goes about daily activities. These constraints motivate smaller models and more efficient algorithms, especially streaming, memory-aware methods that compress and sample long-duration visual inputs smartly. Through this, robust perception, reasoning, and dialogue can run on-device within unified memory limits.

AR system (tracking the gaze real-time, front camera high-resolution, hardware CPU/GPU but has limit compute capacity, this dataset can also be used to train an efficient long-context video processing framework.)

### 2.1 Spatial Dynamics of Human Attention

Human perception is inherently selective, as the brain cannot process the entire visual field with equal fidelity. Prior research has described the attention field as a "spotlight" [19,40], often approximated by gaze location. Attention strength typically decays spatially, commonly approximated by a 2D Gaussian [28], resulting in high fidelity at the focus point that gradually diminishes with distance [10,17].

This attentional pattern strongly influences the types of questions an AR user is likely to ask in real scenarios. As illustrated in Figure 1 (b), a user may focus on a cup while walking in the kitchen. Since attention is concentrated on the cup, the user is more likely to later issue a query about this object (e.g., How many dishes did I wash today?). In contrast, nearby items, shown with lighter bounding boxes in Figure 1 (b), receive weaker attention and are therefore less likely to become the subject of subsequent queries. Similar findings have been reported in cognitive psychology and vision science, where gaze serves as a reliable predictor of future memory recall and questioning behavior [30, 53].

In the field of Psychology and Neuroscience, human attention has historically and empirically been described as a field that decays from a focal point in the visual field. Past research has metaphorized the attention field as a spotlight [19, 40], depicted with a gradient model, and sometimes a Zoom-lens Model [19]. This research all pointed to a 2D Gaussian modeling for the human attention field [28].

Modern AR headsets provide reliable, real-time gaze tracking, which makes it possible to instantiate this 2D Gaussian prior directly on the device. This allows perception and assistance to be aligned with what the user is actually attending to during everyday, long duration use. The gaze-centric attention representation is also computational friendly on embedded AR hardware: it enables foveated sampling, gaze-guided segmentation, and streaming compression of long visual sequences. These features are key tactics when CPU/GPU share limited RAM and power budgets on wearable devices.

### 2.2 Vision Language Models

Contemporary VLMs [6,16,38,45,55] extend language-guided foundation models to additional modalities and demonstrate strong, general capabilities. Trained at scale with spatial data [12], they achieve high accuracy in spatial understanding

and encode rich human preferences and priors. They support perception, reasoning, instruction following, and dialogue, enabling AR assistants that ground semantics in what the user sees and points to.

### 2.3 Long-context Egocentric Video Understanding

Long-context egocentric video understanding has recently gained significant attention in the machine learning community, particularly in extended first-person recordings that capture daily activities and interactions. Such representations enable AR systems to support timely, context-aware assistance by recalling and reasoning over past events in real-world environments. Recent advances in long-context information representation increasingly emphasize structured approaches [5, 7, 8, 15, 47, 51, 52]. In the egocentric vision domain, studies have explored structured video representations by grouping video segments into activity threads [20, 41, 51] or constructing egocentric scene graphs to model object–user relationships [23, 26, 43]. The stored memory entries can later be queried by the user, and these entries are then provided as input to a machine learning model (e.g., VLM) to generate an answer.

Alongside these advances, numerous benchmarks have been introduced to evaluate long-context egocentric video understanding [11, 31, 36, 39, 49, 50, 59]. However, these datasets primarily emphasize generic video-based questions and overlook the human-centric nature of real AR usage. They also restrict questioning to occur only before or after a clip or fixed segment has concluded. In practice, users tend to ask questions anchored to where their attention was directed, often indicated by gaze during recording, yet existing benchmarks fail to capture this critical dimension. Consequently, they fall short of simulating realistic AR scenarios in which attentional focus fundamentally shapes memory retrieval and contextual reasoning. Empirical evidence further supports this view, as incorporating gaze has been shown to significantly improve grounding in egocentric retrieval and natural language query (NLQ) tasks [32].

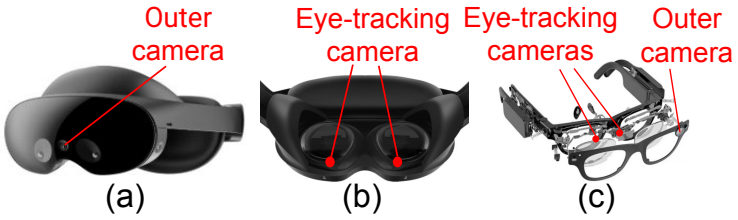


**Fig. 2:** (a) Front and (b) inner views of the Meta Quest Pro headset, which functions as both an AR and VR (Virtual Reality) device. (c) Meta Aria glasses equipped with multiple cameras.

### 2.4 AR System

Figure 2 illustrates typical AR device hardware configurations. These systems feature multiple front- and side-facing cameras that capture the user's field of view and gaze position. Outward-facing cameras generate high-resolution imagery (e.g., $1408 \times 1408$ on the Meta Aria glasses [1]), while inward-facing cameras capture lower-resolution monochrome images of the eyes. Combined, these sensing mechanisms enable gaze tracking [2, 3, 37], typically through either analytical methods such as pupil–corneal reflection modeling or machine learning approaches [33]. These modalities provide critical signals of user attention and interaction, making them foundational for many AR applications.

## 3 Data Collection Procedure

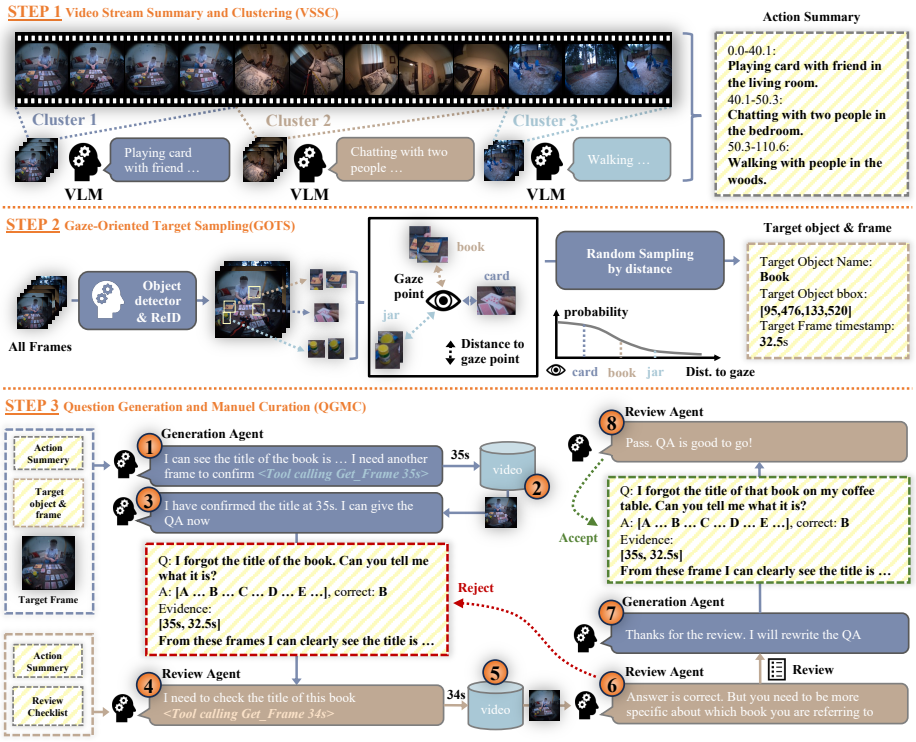


**Fig. 3:** Data generation pipeline for EgoEverything. Step 1, Video Stream Summary and Clustering (VSSC), combines clustering, summary generation, and manual inspection to obtain high-quality video descriptions. Step 2, Gaze-Oriented Target Sampling, detects objects in each frame, computes their distance to gaze, applies ReID to group object instances, and samples target objects by distance. Step 3, Question Generation and Manual Curation, uses a self-feedback Synthesizer Agent and Validator Agent to generate and curate MCQs for each target object.

### 3.1 Overview

The collection process of EgoEverything consists of three steps: (1) *Video Stream Summary and Clustering* (VSSC), (2) *Gaze-Oriented Target Sampling* (GOTS), (3) *Question Generation and Manual Curation* (QGMC). During VSSC, the VLM is prompted with an egocentric video clip to generate a summary of the user's action over the given time span, as shown in Step 1 of Figure 3. Manual inspection is then applied to remove errors in the generated action summaries. Based on the temporal scope of each summary, the egocentric video stream can be categorized accordingly. During GOTS, highlighted in Step 2 in Figure 3, each clustered video tile is examined to sample the objects appearing within it. Sampling is guided by the *Perception Sampler* (PS), which adaptively adjusts its statistical distribution according to the gaze location in the current frame. The selected target objects are then passed to the subsequent stage for question generation. For QGMC, illustrated in Step 3 of Figure 3, we deployed two VLM agents to handle question synthesis and validation. The Synthesizer Agent (VA) creates multiple-choice questions (MCQs) centered on the selected target object, whereas the Validator Agent (VA) examines these questions and delivers feedback. Following VQA generation, the dataset undergoes additional refinement through over **400** hours of human review combined with rule-based screening.

### 3.2 Video Stream Summary and Clustering

EgoEverything is built based on the real-trace egocentric video dataset that included the real gaze tracking trace, including the AriaEveryday Activities (AEA) dataset [34] and Nymeria dataset [35]. The AEA dataset includes 143 daily activity clips across 5 indoor locations, totaling $\sim$ 7.3 hours, while the Nymeria dataset has $\sim$ 300 hours of videos from $\sim$ 50 locations. Both datasets provide egocentric videos and user gaze points recorded by smart glasses.

While the Nymeria dataset provides rich, time-aligned narration text describing major events and interacted objects, the AEA dataset lacks such annotations. To address this gap, we cluster consecutive frames within each video clip and generate action summaries, following Step 1 of Figure 3. Specifically, we first extract frame-level visual features using a pretrained ResNet-50 [24], and then apply the k-means algorithm to group the frames into clusters. The number of clusters $K$ is determined by video duration and scene complexity. For the AEA dataset, we empirically set $K = 12$, noting that larger values of $K$ yield more fine-grained summaries. To mitigate label jitter, we apply temporal smoothing by assigning each frame the most common label within a $\pm$5-frame window. Finally, clips with identical labels are merged into contiguous segments, while enforcing a minimum segment duration of two seconds.

### 3.3 Gaze-Oriented Target Sampling

Using the video segments obtained from VSSC, we introduce the GOTS framework, which simulates human attention mechanism described in Section 2.1 (Step 2 in Figure 3) to questions.

GOTS begins by detecting all objects within each video segment, leveraging a VLM to extract their bounding boxes and labels. Because adjacent frames are often nearly identical, this step generates many redundant detections of the same object over time, which diminishes the diversity of potential questioning targets. To mitigate this, we incorporate a lightweight re-identification (ReID) stage. Specifically, each detected object is cropped and encoded using the visual encoder of a pretrained CLIP model [42]. Detected objects with highly similar CLIP embeddings are then consolidated, ensuring only one representative instance of each object is retained across the sequence.

Subsequently, we measure the Euclidean distance between each object's bounding box centroid and the corresponding gaze position on a per-frame basis. Using this information, we first randomly sample a Target Frame from the video segment, and then sample a single object from the Target Frame as the Target Object based on the PS $S_\theta(\cdot)$. We parameterize $S_\theta(\cdot)$ as a 2D Gaussian distribution in its basic form, expressed as:

$$S_\theta(\cdot) \propto \exp\left(-\frac{\|o-f\|^2}{2\theta^2}\right) \tag{1}$$

Here $f$ denotes the gaze position, and $S_\theta(\cdot)$ gives the selection probability at object centroid $o$. This probability diminishes as the separation $\|o-f\|$ increases, with $\theta$ modulating the decline rate. Multiple target objects are then drawn from this distribution and forwarded to the following question-generation pipeline.

### 3.4 Question Generation and Manual Curation

**Iterative Question Refinement** The Synthesizer Agent (SA) generates questions based on the Target Object, mimicking natural human inquiry patterns. Given the *Target Frame* at time $t_0$, we first sample a questioning timestamp $t_q \in [t_0 + \Delta, T]$, where $T$ is the video end and $\Delta$ is termed *recall interval*. This randomization avoids trivial overlap with the Target Frame while ensuring diverse temporal coverage. Prior work in cognitive science [9, 44, 54] shows that varying the delay between stimulus and questioning can improve memory and comprehension, and that humans flexibly recall events across different temporal spans. Uniform sampling provides a practical and behaviorally plausible baseline for determining when to ask questions.

We employ a pretrained VLM as the Synthesizer Agent (SA), fine-tuned to invoke external tools through specific APIs. To reduce computational costs, SA does not process the full video directly. Instead, it interacts with two APIs: GetFrame, which retrieves a high-resolution frame at a specified timestamp for static detail analysis, and GetSegment, which provides a downsampled video clip over a selected time span for verifying dynamic activities. Guided by the system prompt, SA constructs an MCQ about the Target Object, following Step 3 of Figure 3, with multiple sub-steps. In sub-step 1, SA analyzes the Action Summary and uses the Target Frame to locate the Target Object. The timestamp of the Target Frame provides contextual information about the activity in the

Action Summary, while the associated bounding box for object detection guides SA in identifying the visual features of the Target Object. In sub-step 2, SA first drafts a daily life question about the Target Object, reflecting natural human routines, framed in a natural, human-like style. Then, it identifies the required additional information and iteratively invokes tools and evaluates new evidence until sufficient context is collected to construct an MCQ.

The SA then submits the MCQ with supporting evidence to the Validator Agent (VA) for review (sub-steps 3–4). Similar to the SA, the VA accesses the same Action Summary and may also invoke tools to inspect portions of the video during its evaluation. Unlike the SA, the VA does not receive the Target Object or Target Frame. It verifies factual accuracy, identifies ambiguities, evaluates question clarity, and provides at least one additional piece of evidence to enhance the credibility of the MCQ (sub-step 5). The VA returns feedback to the SA (sub-step 6), which refines and resubmits the MCQ to the VA. If all checks pass, the VA finalizes the MCQ. MCQs failing after two review rounds are discarded.

**Manual Filtering and Labeling** After question generation, we obtain high-quality MCQs; however, certain failure modes may still lead to low-quality outputs. The most common case arises when the Target Object in the Target Frame is ambiguous due to factors such as distance, occlusion, inadequate lighting, or viewpoint distortion. In such cases, the object detector may misclassify the Target Object as another item. A second failure mode arises when the Agent misinterprets spatial layouts, resulting in view-dependent or incorrect spatial descriptions. Finally, certain target objects are inherently unsuitable for MCQ generation, producing questions that do not match typical AR user queries.

To mitigate these issues, we first apply rule-based filtering to exclude target objects that are unsuitable for questioning (e.g., walls, ceilings, floors, or the camera wearer's body parts). We further discard MCQs that violate typical AR user questioning patterns, such as those referencing timestamps or explicitly mentioning the word "video." Next, we conduct human review. Annotators are presented with each MCQ together with the corresponding video. Without access to the correct answer, they are asked to select one choice from five options. If minor issues are observed, annotators may refine the MCQ; for major flaws, they mark the item as invalid. After the review process, we retain only those MCQs whose pseudo answers are consistent with the annotators' selections.

Finally, we apply large language model (LLM)-based blind filtering: the LLM is prompted to guess the correct answer without access to the video, and we retain only those MCQs it answers incorrectly. This ensures that the retained questions cannot be solved by simple logical reasoning or textual cues alone, preventing video-free answering in LEU.

## 4 Evaluation

EgoEverything is developed on the foundation of real-trace egocentric video datasets enriched with authentic gaze tracking information, incorporating data

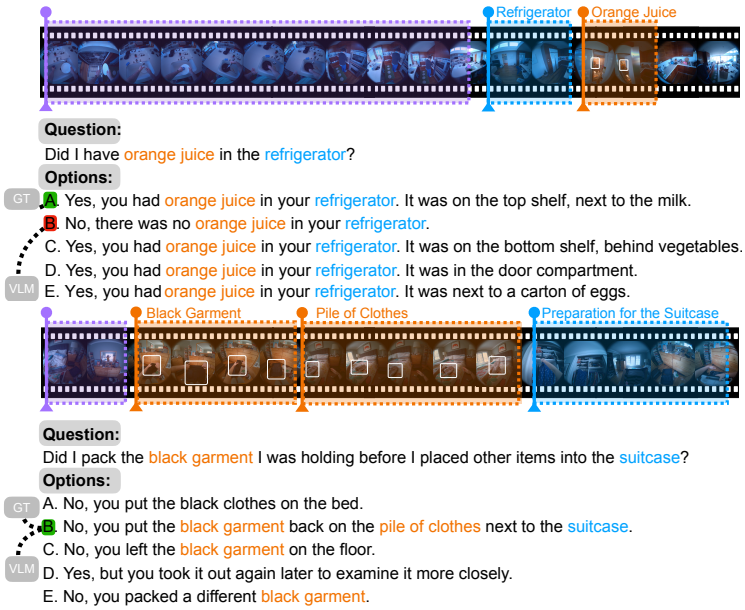


**Fig. 4:** Video frames and keywords in the MCQ are highlighted in different colors. Orange indicates the target objects, Blue highlights contextual entities mentioned in the question or answer, and Purple denotes regions where the target objects cannot be located. The GT label refers to the ground truth option, and the VLM label refers to the option selected by the model.

from the the Aria Everyday Activities [34] (AEA) and Nymeria [35] datasets. The AEA dataset [34] comprises 143 egocentric video clips captured at $1408 \times 1408$ resolution (averaging 41.6 MB per minute), spanning five indoor environments with various daily activities across roughly 7.3 hours of footage. Each video is paired with synchronized gaze traces from wearable AR devices, making AEA a compact yet carefully annotated benchmark well-suited for fine-grained studies of attention in everyday tasks. In contrast, the Nymeria dataset offers a much larger-scale resource, with videos recorded at $2016 \times 2208$ resolution (averaging about 49.3 MB per minute) and approximately 100 hours of recordings across nearly 50 diverse indoor and outdoor locations. These videos include real gaze traces along with naturalistic variations in activity, environment, and lighting, providing a rich foundation for training and evaluating models on long-duration and heterogeneous egocentric experiences. Together, AEA and Nymeria complement each other by combining curated activity-focused data with large-scale diverse traces, enabling comprehensive exploration of egocentric understanding. Building on these resources, EgoEverything contains over 5,000 multiple-choice question–answer pairs spanning more than 100 hours of video. Dataset examples are illustrated in Figure 4.

In the generation stage, we use a 2D Gaussian distribution as the PS, shown in Equation 1, with the standard deviation $\theta = 400$ pixels and recall interval $\Delta = 3$ minutes. For annotation, we developed a web-based labeling tool with 12 trained annotators, who collectively labeled about 21,600 questions over 400 hours. The MCQ adoption rate during annotation was approximately 70%, while rule-based and blind filtering yielded acceptance rates of around 50%.

We evaluate several recent VLMs on EgoEverything, including Videollama3 [56], Gemini [46], LongVA [57], and Llava-Video [58]. Since our MCQs are sampled using a Gaussian-based PS, content near the gaze location tends to be more relevant for solving MCQs than distant information.

To validate this property, we design several preprocessing baselines: *Gaze Crop* (GC), crops each frame around the gaze fixation location using a square bounding box covering roughly 10% of the original frame size; *Gaze Mask* (GM), retains the complementary regions outside GC; *Average Downsampling* (AD), uniformly downsamples each frame to 10% of its original resolution; and *Full Resolution* (FR) uses original frames. In addition, we evaluate two recent methods for LEU tasks, AMEGO [23] and VideoMindPalace (VMP) [26], which construct egocentric scene graphs to capture key object–user relationships while filtering redundant information, achieving strong performance on standard LEU benchmarks such as EgoSchema [36] and NExT-QA [50]. We apply AMEGO and VMP over Videollama3 and Gemini. Our goal is to examine how these methods perform on this real-life, human attention-driven LEU dataset.

*Question Categories* For question generation, we manually classified the questions into eight distinct categories (Figure 5): **Item Presence**, what item appears in someplace. **Appearance**, what color/shape something is. **Event Verification**, did something happen. **State Verification**, what is the state of something. **Spatial–Spatial**, ask relative location between objects. **Direct Location**, ask absolute location of something. **Temporal–Spatial**, ask where something is when something happens. and **Others**.

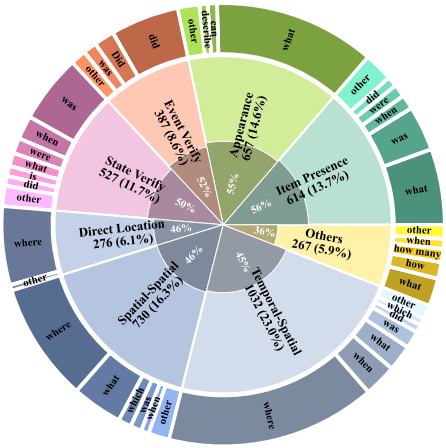


**Fig. 5:** The middle ring shows the proportion of each category and the number of MCQs in it. The outer ring shows the most frequent interrogative words in that category. The inner circle shows the average accuracy of all VLM models from Table 1 for each category.

We manually classified the generated questions into eight categories: **Item Presence**: asking whether an item appears in a given place. **Appearance**: asking about the color, shape, or other visual attributes of an item. **Event Verification**: asking whether a specific event occurred. **State Verification**: asking about the state or condition of an item. **Spatial–Spatial**: asking about the relative location between objects.

**Direct Location**: asking about the absolute location of an item. **Temporal–Spatial**: asking where an item is when another event occurs. **Others**: miscellaneous questions that do not fit the above categories. Figure 5 provides a comprehensive overview of the distribution across question categories.

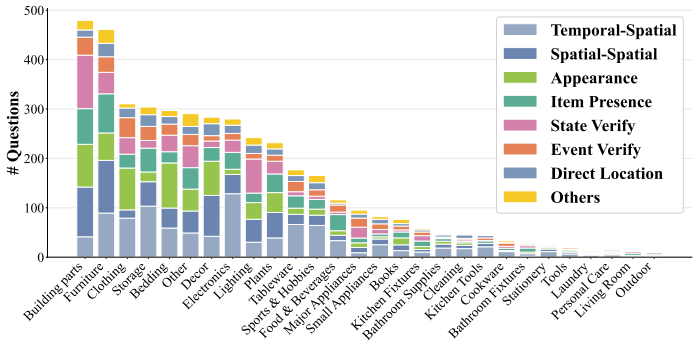


**Fig. 6:** Distribution of Target Object Categories. The x-axis represents object categories, the y-axis counts the number of MCQs generated using the target object in a given category. The length of each colored bar segment indicates the number of questions of that type, revealing the reasoning associated with each target object.

*Target Object Category:* As described in Section 3.3, our GOTS framework selects Target Objects when generating MCQs. Selected Target Objects are divided into 28 categories and show the distribution in Figure 6, where the y-axis denotes the number of associated MCQs and the stacked colors represent the proportions of different question categories within each Target Object category.

### 4.1 Accuracy Evaluation on EgoEverything

As shown in Table 1, among the VLMs, under the full-resolution setting where the entire input frame is provided for processing, Gemini achieves the highest accuracy of 63.1% across the MCQs, while the other models perform even lower. Nevertheless, all VLMs remain far behind human performance, as our human annotators reach an average accuracy of 83.5% across 12 participants. This gap highlights the clear performance deficiency of current VLMs on EgoEverything.

**Table 1:** Performance comparison of different methods across VLMs on EgoEverything. "NA" means not available. The human annotators can achieve an average accuracy of 83.5%.

| Model | FR | AD | GC | GM | VMP | AMEGO |
|---|---|---|---|---|---|---|
| Videollama3-7b [56] | 49.1 | 46.1 | 42.4 | 35.2 | 20.2 | 19.5 |
| Videollama3-2b | 46.5 | 44.9 | 40.5 | 34.5 | 21.3 | 19.4 |
| Gemini 1.5 pro [46] | 63.1 | 58.4 | 52.7 | 37.7 | 33.2 | 18.3 |
| LongVA [57] | 34.6 | 31.6 | 28.9 | 22.2 | NA | NA |
| Llava-Video [56] | 42.6 | 36.6 | 32.9 | 26.0 | NA | NA |

All input processing methods, including GC, GM, AD, VMP, and AMEGO, present a degradation in MCQ prediction accuracy compared to FR. Among

them, GC performs better than GM because the MCQs are generated according to the PS centered at the gaze fixation location, and this method preserves the most critical information. In contrast, GM suffers a substantial performance drop since it excludes these key details regarding human attention specified by the gaze fixation location. Interestingly, AD outperforms GC by preserving not only the gaze-centered visual content but also peripheral information that remains important for answering MCQs. Finally, both VMP and AMEGO achieve the lowest performance among the evaluated methods. This is because neither approach infers directly from raw video. Instead, they convert the video into structured text via object detection and use LLM for pure text reasoning. Both methods only detect interacted objects and overlook many activity-irrelevant objects in our benchmark.

| Model | Videollama3-7b | Gemini 1.5 pro | LongVA |
|---|---|---|---|
| Accuracy (%) | 22.9 | 35.9 | 21.8 |

**Table 2:** VLM model performance under blind setting.

In addition, to test whether the MCQs within EgoEverything can be solved by solely inspecting the textual information from the question, we conduct a text-only evaluation. This setting is undesirable because it does not allow the model to leverage visual information during processing, and thus may encourage reliance on linguistic shortcuts instead of true multimodal reasoning. To achieve this, only the textual input within each data of MCQ is delivered to VLM and the video frames are eliminated. As indicated by Table 2, this will greatly degrade the accuracy for Videollama3-7b [56], Gemini [46] and LongVA [57], confirming that our questions cannot be answered from text alone.

As described in Section 3.4, when generating MCQs we mimic real-life AR scenarios by introducing randomized questioning times $t_q$, whereas in other LEU datasets questioning typically occurs only after a clip or fixed segment has ended. To study how variation in $t_q$ impacts accuracy, we select the 10% of MCQs with the largest recall intervals ($\Delta$) and the 10% with the smallest intervals, and then measure the accuracy of Videollama3-7B over EgoEverything. As shown in Table 3 (c), the results reveal that accuracy drops markedly as the recall interval increases, from 49% to 33.3%. This indicates that current models exhibit substantial performance inconsistency under human-like diverse questioning times, demonstrating that variation in questioning time directly impacts LEU accuracy and thereby validating our randomized questioning-time design.

*Impact of Object-Gaze Distance* During generation, we sample a target object according to its distance from the gaze point $\|o-f\|$, as defined in Equation 1. This better simulates human attention and produces questions that align with user focus. To examine the impact of object–gaze distance, we report Videollama3-7b's accuracy on our MCQs grouped by $\|o-f\|$, as shown in Figure 3 (a). The results showed that Target Objects located at the periphery of the visual field produce more challenging MCQs than Target Objects near the center of attention, with accuracy decreasing from 55% to 30% as distance increased. These results indicate that current models struggle with MCQs related to peripheral

information. Unlike prior benchmarks, our attention-aware generation produces both challenging peripheral questions and user-focused questions.

*Impact of Object Size* To examine whether VLMs attend to objects of different scales in a comparable manner, we measure bounding box sizes and analyze accuracy as a function of target object area (in pixels$^2$). Area thresholds ranging from $1 \times 10^0$ to $1 \times 10^5$ pixels$^2$ were applied. As shown in Figure 3 (b), accuracy consistently increases with larger bounding box thresholds. Across the evaluated models, performance is biased toward larger objects, while smaller and less salient objects are frequently overlooked.

In summary, these results expose systematic limitations: the evaluated VLMs perform worse on objects that are farther from the gaze point, require longer recall intervals, or appear at smaller scales. As a result, current models still lack the robustness needed for reliable everyday assistance in AR scenarios.

**Table 3:** Accuracy analysis under three factors. Dist. denotes $\|o - f\|$ (pixel), area denotes bounding box area (pixel$^2$), and recall interval denotes $t_q - t_0$.

| **(a) Dist. Gaze** | | **(b) BBox area** | |
|---|---|---|---|
| Dist. (pixel) | Acc. (%) | Area (pixel$^2$) | Acc. (%) |
| 0 | 54.3 | 100 | 44.3 |
| 50 | 54.2 | 500 | 44.4 |
| 100 | 54.2 | 2k | 44.5 |
| 150 | 54.9 | 5k | 44.9 |
| 200 | 54.8 | 10k | 49.2 |
| 300 | 51.8 | 50k | 54.8 |
| 400 | 47.2 | 100k | 54.8 |

| **(c) Recall interval** | |
|---|---|
| Percentile | Acc. (%) |
| top 10% | 33.3 |
| bottom 10% | 49.0 |

## 5 Conclusion

We introduced EgoEverything, a benchmark for long-context egocentric video understanding that explicitly models human attention in AR settings. EgoEverything leverages an attention-inspired target sampling strategy (guided by real gaze) and a multi-agent VQA generation pipeline, followed by rule-based screening, blind filtering, and extensive human review, to produce over 5,000 high-quality multiple-choice questions spanning more than 100 hours of egocentric video. Our evaluation across several state-of-the-art VLMs reveals a substantial gap to human performance and consistent failures on attention-driven and long-recall questions, highlighting that current models still struggle to reason over long temporal contexts under realistic, user-centric querying patterns. We hope EgoEverything will serve as a practical testbed for robust, attention-aware long-context video understanding in AR assistants.

## References


1. Meta aria glasses, https://www.projectaria.com/glasses/
2. Meta quest 3., https://www.meta.com/quest/quest-3/
3. Meta quest pro, https://www.meta.com/quest/quest-pro/
4. Al-Ansi, A.M., Jaboob, M., Garad, A., Al-Ansi, A.: Analyzing augmented reality (ar) and virtual reality (vr) recent development in education. Social Sciences & Humanities Open **8**(1), 100532 (2023)
5. Arnab, A., Sun, C., Schmid, C.: Unified graph structured models for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8117–8126 (2021)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
7. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 105–121 (2018)
8. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: 2011 International Conference on Computer Vision. pp. 778–785. IEEE (2011)
9. Carpenter, S.K., DeLosh, E.L.: Application of the testing and spacing effects to name learning. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition **19**(5), 619–636 (2005)
10. Carrasco, M.: Visual attention: The past 25 years. Vision research **51**(13), 1484–1525 (2011)
11. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Fei-Fei, L.: Hourvideo: 1-hour video-language understanding (2024), https://arxiv.org/abs/2411.04998
12. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities (2024), https://arxiv.org/abs/2401.12168
13. Chengoden, R., Victor, N., Huynh-The, T., Yenduri, G., Jhaveri, R.H., Alazab, M., Bhattacharya, S., Hegde, P., Maddikunta, P.K.R., Gadekallu, T.R.: Metaverse for healthcare: a survey on potential applications, challenges and future directions. IEEE Access **11**, 12765–12795 (2023)
14. Chidsin, W., Gu, Y., Goncharenko, I.: Ar-based navigation using rgb-d camera and hybrid map. Sustainability **13**(10), 5585 (2021)
15. Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., Yang, M.Y.: Spatial-temporal transformer for dynamic scene graph generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16372–16382 (2021)
16. Deitke, M., et al., C.C.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models (2024), https://arxiv.org/abs/2409.17146
17. Desimone, R., Duncan, J., et al.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**(1), 193–222 (1995)
18. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)
19. Eriksen, C.W., St. James, J.D.: Visual attention within and around the field of focal attention: A zoom lens model. Perception & Psychophysics **40**(4), 225–240 (1986). https://doi.org/10.3758/BF03211502, https://doi.org/10.3758/BF03211502

20. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)
21. Gerup, J., Soerensen, C.B., Dieckmann, P.: Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review. International journal of medical education **11**, 1 (2020)
22. Goesele, M., Andersen, D., Chen, Y., Green, S., Ilg, E., Li, C., Liu, J., Kuo, G., Wan, L., Newcombe, R.: Imaging for all-day wearable smart glasses (2025), https://arxiv.org/abs/2504.13060
23. Goletto, G., Nagarajan, T., Averta, G., Damen, D.: Amego: Active memory from long egocentric videos. In: European Conference on Computer Vision. pp. 92–110. Springer (2024)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hou, B.J., Abdrabou, Y., Weidner, F., Gellersen, H.: Unveiling variations: A comparative study of vr headsets regarding eye tracking volume, gaze accuracy, and precision. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 650–655. IEEE (2024)
26. Huang, Z., Ji, Y., Wang, X., Mehta, N., Xiao, T., Lee, D., Vanvalkenburgh, S., Zha, S., Lai, B., Yu, L., et al.: Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24169–24179 (2025)
27. Inc., M.P.: Meta quest 3. https://www.meta.com/quest/quest-3/ (2023)
28. Ioannides, A., Poghosyan, V.: The early spread of spatial and non-spatial attentional effects in human visual cortex. In: Proceedings of the Frontiers in Neuroscience Conference. vol. 4, pp. – (Mar 2010). https://doi.org/10.3389/conf.fnins.2010.06.00379, https://doi.org/10.3389/conf.fnins.2010.06.00379
29. Jo, Y.J., Choi, J.S., Kim, J., Kim, H.J., Moon, S.Y.: Virtual reality (vr) simulation and augmented reality (ar) navigation in orthognathic surgery: a case report. Applied Sciences **11**(12), 5673 (2021)
30. Land, M.F., Hayhoe, M.: In what ways do eye movements contribute to everyday activities? Vision research **41**(25-26), 3559–3565 (2001)
31. Li, J., Wei, P., Han, W., Fan, L.: Intentqa: Context-aware video intent reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11963–11974 (2023)
32. Lin, W.C., Lien, C.M., Lo, C., Yeh, C.H.: Gazenlq @ ego4d natural language queries challenge 2025 (2025), https://arxiv.org/abs/2506.05782
33. Liu, W., Duinkharjav, B., Sun, Q., Zhang, S.Q.: Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality. IEEE Transactions on Visualization and Computer Graphics (2025)
34. Lv, Z., Charron, N., Moulon, P., Gamino, A., Peng, C., Sweeney, C., Miller, E., Tang, H., Meissner, J., Dong, J., et al.: Aria everyday activities dataset. arXiv preprint arXiv:2402.13349 (2024)
35. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024)

36. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
37. Microsoft: HoloLens 2 Specs (2023), https://learn.microsoft.com/en-us/hololens/hololens2-hardware
38. OpenAI, et al., J.A.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
39. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., Chalk, J., Zhu, Z., Guerrier, R., Abdelazim, F., Zhu, B., Moltisanti, D., Wray, M., Doughty, H., Damen, D.: Hd-epic: A highly-detailed egocentric video dataset (2025), https://arxiv.org/abs/2502.04144
40. Posner, M.: Orienting of attention. Q J Exp Psychol **32**, 3–25 (01 1980)
41. Price, W., Vondrick, C., Damen, D.: Unweavenet: Unweaving activity stories. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13770–13779 (2022)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
43. Rodin, I., Furnari, A., Min, K., Tripathi, S., Farinella, G.M.: Action scene graphs for long-form understanding of egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18622–18632 (2024)
44. Roediger III, H.L., Karpicke, J.D.: Test-enhanced learning: Taking memory tests improves long-term retention. Psychological science **17**(3), 249–255 (2006)
45. Team, G., et al., P.G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://arxiv.org/abs/2403.05530
46. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
47. Wang, Y., Yang, Y., Ren, M.: Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. arXiv preprint arXiv:2312.05269 (2023)
48. Westin, T., Neves, J., Mozelius, P., Sousa, C., Mantovan, L.: Inclusive ar-games for education of deaf children: Challenges and opportunities. In: European Conference on Games Based Learning. vol. 16, pp. 597–604 (2022)
49. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
50. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
51. Yang, Y., Ren, M.: Memory storyboard: Leveraging temporal segmentation for streaming self-supervised learning from egocentric videos. arXiv preprint arXiv:2501.12254 (2025)
52. Yang, Y., Zhao, Z., Shukla, S.N., Singh, A., Mishra, S.K., Zhang, L., Ren, M.: Streammem: Query-agnostic kv cache memory for streaming video understanding. arXiv preprint arXiv:2508.15717 (2025)
53. Yarbus, A.L.: Eye movements and vision. Springer (2013)
54. Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: a mind-brain perspective. Psychological bulletin **133**(2), 273 (2007)

55. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding (2025), https://arxiv.org/abs/2501.13106
56. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
57. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024), https://arxiv.org/abs/2406.16852
58. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
59. Zhou, W., Cao, K., Zheng, H., Zheng, X., Liu, M., Kristensson, P.O., Mayol-Cuevas, W., Zhang, F., Lin, W., Shen, J.: X-lebench: A benchmark for extremely long egocentric video understanding. arXiv preprint arXiv:2501.06835 (2025)