EGOQUESTIONS:CRAFTING EGOCENTRIC QUESTIONS FOR EGOCENTRIC VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

A thorough understanding of models' egocentric capabilities is crucial for robotics, autonomous driving, smart glasses, etc. Egocentric VideoQA aims to assess models' understanding of first-person videos, but existing benchmarks often include questions that do not reliably probe recorder-centric reasoning. Using these datasets to train and evaluate models can obscure true model capabilities and reduce the value of curated egocentric data. To address this, we define egocentric questions and propose three clear principles: a question should focus on the video recorder and their activities; it must avoid shortcut cues that allow answers via generic scene or action recognition (e.g., simultaneously naming an action and its object); while intentions and attributes may serve as shortcuts for actions and objects, those that require understanding of the recorder's perspective will not. Guided by these principles, we build a checking pipeline to filter existing QA pairs and a crafting pipeline to generate valid egocentric questions. We release EgoQuestions, a benchmark of 2,500 curated egocentric QA instances created with our pipeline, and evaluate several proprietary and open-source VLMs. Results reveal substantial room for improvement in current models' egocentric capabilities and a clear performance gap (about **10**%) between egocentric questions that adhere to our principles and flawed alternatives, demonstrating existing egocentric benchmarks tend to overrate models' first-person capabilities. and the need for rigorously designed egocentric benchmarks to more accurately assess models' first-person vision capabilities.

1 INTRODUCTION

Egocentric videos refer to recordings captured from a first-person point of view, which have become popular with the advancements of wearable technology (Hodges et al., 2006; Engel et al., 2023). The field of egocentric vision has gained significant attention from researchers in recent years, with applications across various domains, including augmented reality (AR) (Plizzari et al., 2024; Lv et al., 2024), embodied AI (Mu et al., 2023), and human-object interaction (Liu et al., 2022; Wang et al., 2023) among others. To support the development of models for understanding and describing these videos, numerous datasets for egocentric video analysis have emerged (Damen et al., 2018; Grauman et al., 2022; Zhu et al., 2023; Pan et al., 2023; Grauman et al., 2024; Bi et al., 2024; Perrett et al., 2025; Chen et al., 2025), including datasets such as Ego4D and Ego-Exo4D that provide extensive collections of videos. Egocentric video understanding presents additional challenges compared to videos captured from a third-person (exocentric) perspective. The scene of egocentric videos changes constantly and unpredictably, and usually only the recorder's some body parts are visible. More importantly, for effective real-world deployment, a model must be able to perceive its environment from a first-person viewpoint. However, Vision-Language Models (VLMs) trained predominantly on third-person visual data may not possess this essential egocentric capability.

A variety of benchmarks have been proposed to evaluate models' egocentric capabilities, including those that adapt Video Question Answering (VideoQA) to egocentric videos (Fan, 2019; Jia et al., 2022; Bärmann & Waibel, 2022; Mangalam et al., 2023; Cheng et al., 2024b; Di & Xie, 2024; Cheng et al., 2024a; Zhou et al., 2025a; Chen et al., 2025). While these benchmarks encompass egocentric videos and specific tasks, they fail to align the videos with corresponding genuinely egocentric questions, i.e., the questions regarding the video recorder's intention, action, and perception of the

Figure 1: Instances in EgoTextVQA (Zhou et al., 2025a), EgoTempo (Plizzari et al., 2025), QaEgo4D (Bärmann & Waibel, 2022) and EgoQuestions. We identify shortcomings in the questions contained in the three existing benchmarks and describe them respectively in the image above.

scene. For instance, as shown in Figure 1, some questions (top-left) focus on general scene content, which is unrelated to the recorder's perspective. Others (top-right) allow models to rely on non-egocentric cues like spatial grounding or simple action recognition, bypassing the need for genuine egocentric reasoning. Furthermore, some questions are poorly constructed, such as those where the answer is implied within the question itself (bottom-left). These question flaws may misleadingly suggest that current models have a strong egocentric understanding while diminishing the value of curated egocentric video data. We believe that evaluating a model's first-person capabilities requires not only first-person perspective videos but also carefully crafted **questions** tailored to these scenes.

To address defects in existing questions, in this paper, we aim to craft **egocentric questions** tailored for **egocentric tasks**. We first define egocentric questions and state three principles they must follow. We propose that the question must focus on the camera wearer, not merely on the environment. Also, the question must avoid shortcuts that let models rely solely on general scene or action recognition—for example, it should not simultaneously name the action and the object that makes the answer obvious. Finally, the question may ask about the wearer's intentions or attributes when those aspects require understanding the recorder's role. Based on these deterministic principles, we develop a checking pipeline to evaluate existing question–answer pairs and a crafting pipeline to generate valid egocentric questions.

We introduce the EgoQuestions benchmark, which includes 2,500 instances of egocentric questions paired with curated egocentric video clips, all generated using our crafting pipeline. We evaluate several recent VLMs, including proprietary and open-source models. Results show substantial room for improvement in current VLMs' egocentric understanding. To illustrate the effect of flawed questions, we also construct corresponding question sets that violate our principles; performance on these flawed questions is substantially different from performance on our egocentric questions, underscoring the need to use proper egocentric questions in egocentric VideoQA.

Our contributions can be summarized in three key abstracts.

We identify current shortcomings in questions and propose a set of principles for egocentric questions. Ours is the first work to identify defects in questions in current egocentric VideoQA benchmarks. Furthermore, we propose explicit requirements that egocentric questions must satisfy and provide principles for validating them.

We develop a benchmark, EgoQuestions, that encompasses crafted egocentric questions. We curate the questions of EgoQuestions with our question-crafting pipeline, producing questions of substantially higher quality than those in current benchmarks. We further employ this benchmark to evaluate recent VLMs, reporting results that reflect the models' egocentric capabilities. We also

demonstrate the performance gap caused by problematic questions by constructing a comparison set of flawed questions, which indicates that existing egocentric benchmarks tend to overestimate models' first-person capabilities.

We implement a crafting pipeline to generate egocentric questions. We will release our full crafting pipeline for egocentric question generation, enabling the community to generate more egocentric questions and facilitate the development of VLMs.

2 Related Work

2.1 Egocentric Video Question Answering Benchmarks

Video question answering is a task that requires a joint understanding of both visual content and the corresponding textual information. With the growing accessibility of wearable photographic devices and the emergence of large-scale egocentric video datasets (Grauman et al., 2022; Wang et al., 2023; Grauman et al., 2024; Perrett et al., 2025), previous studies have investigated the integration of textual cues with egocentric videos, leading to the development of egocentric VideoQA benchmarks. Among these, EgoTempo (Plizzari et al., 2025) emphasizes temporal comprehension of videos, whereas EgoTextVQA (Zhou et al., 2025a) is built to evaluate the scene-text understanding capabilities of models. Additional addressed challenges include long-form video orientation (Mangalam et al., 2023; Zhou et al., 2025b), goal understanding (Jia et al., 2022), social norm interpretation (Rezaei et al., 2025), multi-hop video question answering (Chen et al., 2025), cross-domain generalization (Li et al., 2025), *etc*.

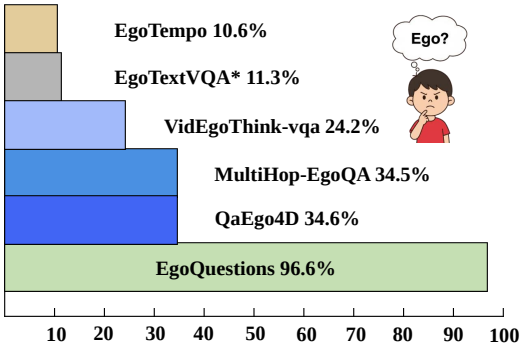

Figure 2: A visualization of the percentage of egocentric questions contained in current egocentric VideoQA benchmarks and EgoQuestions. * represents that we only evaluate a subset of EgoTextVQA.

While defining tasks to utilize egocentric videos, prior works have largely overlooked the linguistic formulation of the corresponding questions. The percentage of egocentric questions contained in current benchmarks are depicted in Figure 2. This may lead to emergence of inferential shortcuts that models can identify, resulting in a gap between evaluation outcomes and the model's true egocentric capabilities. In our work, we introduce a definition of egocentric questions accompanied by a framework for generating such questions from narrative descriptions, thereby mitigating this possible discrepancy. As shown in the figure, while other benchmarks contain less than 40% instances with egocentric questions, EgoQuestions boasts a percentage of over 95%.

2.2 Annotations Methods involving LLMs / MLLMs

The recent development of large language models (LLMs) and multi-modal large language models (MLLMs) (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Team et al., 2023; Grattafiori et al., 2024; OpenAI et al., 2024; OpenAI, 2024; DeepSeek-AI et al., 2025) has enabled the usage of these models for benchmark construction. Among benchmarks concerning Egocentric VideoQA, several prior studies have employed LLMs / MLLMs in their data generation pipelines, primarily leveraging the extensive annotations available in large-scale egocentric video datasets such as Ego4D (Grauman et al., 2022). EgoSchema (Mangalam et al., 2023) utilizes LLMs in both the QA generation and filtering stages, while EgoTextVQA (Zhou et al., 2025a) and VidEgoThink (Cheng et al., 2024a) integrate the usage of the GPT-4o model (OpenAI, 2024) in their QA generation processes.

However, even when provided with detailed prompts, LLMs and MLLMs often struggle to meet complex requirements, necessitating human filtering after LLM / MLLM usage. In our work, we

utilize code templates to systematically generate questions and answers, leveraging LLMs exclusively for grammar component extraction and question refinement. This methodology enables more fine-grained control over the format of questions.

3 What Makes a Good Egocentric Question

In this section, we introduce the defects present in question formulations of current egocentric benchmarks. Subsequently, we propose three guiding principles for the design of egocentric questions and provide an evaluation of existing egocentric VideoQA benchmarks in light of these principles.

3.1 Defects in current questions

The objective of advancing model comprehension of egocentric videos is to facilitate model development in real-world applications, such as personalized agents and embodied AI. Therefore, egocentric VideoQA benchmarks should be designed to assess models' capacity for understanding surrounding environments and interactions from a first-person perspective. However, an examination of current egocentric VideoQA benchmarks reveal that many questions fail to meet this objective, either by not addressing the subject at all or by offering inference shortcuts that do not require egocentric understanding. For instance, the question "What is the man in red clothes doing?" does not concern the video recorder (*i.e.* myself), making it equally applicable to third-person VideoQA tasks. More subtle flaws arise in questions like "What am I cutting that is on the table?". A model proficient in scene identification and action recognition could answer this by detecting the action ("cut") and object location ("on the table") without egocentric comprehension. These deficiencies may result in a misalignment between a model's egocentric capabilities and its performance on related benchmarks.

3.2 Three Principles of Egocentric Questions

To systematically assess existing questions and create new egocentric questions, we propose three principles that such questions should follow.

Principle 1: The question must concern the recorder. To ensure models adopt the recorder's perspective, questions must pertain to the recorder themselves. A question unrelated to the recorder could be equally answered using third-person (exocentric) videos, reducing egocentric videos to conventional exocentric videos with frequent camera movement. For example, "What am I doing?" is an egocentric question, while "What is the man in red clothes doing?" is not.

Principle 2: Avoid shortcuts related to both the action and the object. With the subject restricted to the recorder by Principle 1, there must not be shortcuts related to both the action and the object being interacted with. Otherwise, an MLLM capable of interpreting actions and scenes could answer the question without egocentric understanding. For objects, stating the object itself or its attributes (independent of the recorder) can create shortcuts. A similar logic applies to actions and their associated adverbs. For example, "What am I cutting?" is an egocentric question, but "What am I cutting that is on the table?" is not. Further details are discussed in the following principle.

Principle 3: Attributes or intentions relevant to the recorder are allowed. This principle complements Principle 2. While some attributes may introduce inference shortcuts, those directly relevant to the recorder are unlikely to do so. The model must still understand the recorder's role to comprehend them. The same reasoning applies to the intentions of the actions.

We observe that the aforementioned principles can be reformulated as rules concerning the grammatical components of both simple and complex sentences. Specifically, assuming the question and answer can be restructured into a simple or compound sentence, Principle 1 requires the subject to be the recorder; Principle 2 concerns the appearance of the predicate, object, attributives, and adverbials in the question; and Principle 3 focuses on attributives and adverbials. The relationship helps determine whether a question is truly egocentric.

3.3 Reviewing current benchmarks

Building upon the above three principles, we implement a framework that utilizes LLMs to identify actions, objects, and other components in questions and answers, and to determine whether

Table 1: Evaluation results of questions in QaEgo4D (Bärmann & Waibel, 2022), the VQA task in VidEgoThink (Cheng et al., 2024a), EgoTextVQA (Zhou et al., 2025a), EgoTempo (Plizzari et al., 2025), and MultiHop-EgoQA (Chen et al., 2025) with our framework. * indicates that we only evaluate a certain subset of the items, with details contained in the appendix.

Benchmark	Video Source	Q-A Source	Percentage(%)
QaEgo4D	Ego4D (Grauman et al., 2022)	Ego4D NLQ	34.59
VidEgoThink-vqa	Ego4D	Ego4D Narrations	24.17
EgoTextVQA*	RoadTextVQA, EgoSchema	Videos	11.34
EgoTempo	Ego4D	Ego4D Narrations	10.6
MultiHop-EgoQA	Ego4D	Ego4D Narrations	34.54

a question conforms to our principles. We apply this framework to five benchmarks containing open-ended egocentric VideoQA instances; the percentages of the questions adhering to the three principles, along with each benchmark's video and QA sources, are shown in Table 1. We exclude EgoSchema(Mangalam et al., 2023) and the HD-Epic VQA benchmark(Perrett et al., 2025) in this evaluation because their multiple-choice instances can depend heavily on the provided options. For example, an EgoSchema item asks the model to "briefly describe this interruption and its significance within the video," with detailed descriptions given among the choices. The results show that many existing works overlook the importance of well-structured questions and do not follow our proposed principles. Improving question formatting and adherence to these principles could increase the quality and reliability of responses in VideoQA tasks.

4 EgoQuestions Dataset

In this section, we provide an overview of the construction process of EgoQuestions. We introduce the video collection methods and describe the pipeline used for generating question-answer pairs.

4.1 Video Collection

Our video data are sourced from two recognized and publicly accessible egocentric video datasets: Ego4D(Grauman et al., 2022) and HD-Epic(Perrett et al., 2025). Both datasets contain a diverse variety of manually collected videos. Ego4D includes more than 3,670 hours of egocentric videos, recorded in 74 locations across 9 different countries. HD-Epic's annotators recorded their kitchen activities over at least three days, producing egocentric videos that capture a wide array of objects and actions within complex indoor environments.

The natural language queries (NLQ) task, part of the Episodic Memory benchmark of Ego4D, provides samples with a natural language query Q and a time slot r that serves as the ground truth for when the answer to Q occurs. QaEgo4D (Bärmann & Waibel, 2022) was built upon NLQ annotations, employing human annotators to curate answers to Q. These responses form complete items consisting of a video V, a question Q, a time slot τ, and a newly curated answer A. In our project, we directly use the V and τ from QaEgo4D items for EgoQuestions. Additionally, HD-Epic's diverse video content provides broad coverage of indoor activities, which complements the international and varied settings presented in Ego4D, enhancing the robustness of our dataset.

The second method of utilizing Ego4D videos is based on the dataset's original narrations, each of which includes a timestamp and a brief description of the camera wearer's activities at that moment. Previous studies (Lin et al., 2022; Plizzari et al., 2025) have constructed time slots based on these narrations. We detail their methodologies and present our analysis in the appendix. In our approach, we first sort all the narrations for a video according to their timestamps. We then define the time slots using the following strategy:

$$\tau_k = (\text{random}(t_{k-1}, t_k), \quad \text{random}(t_k, t_{k+1})), \tag{1}$$

where τ_k represents the constructed time slot for the kth narration, and t_k is the original timestamp of that narration. By employing this approach, we account for the density of narrations; in particular, the duration of each time slot is extended based on the spacing between narrations. This helps mitigate the influence of isolated narrations and reduces distractions caused by adjacent narration timestamps within our time intervals.

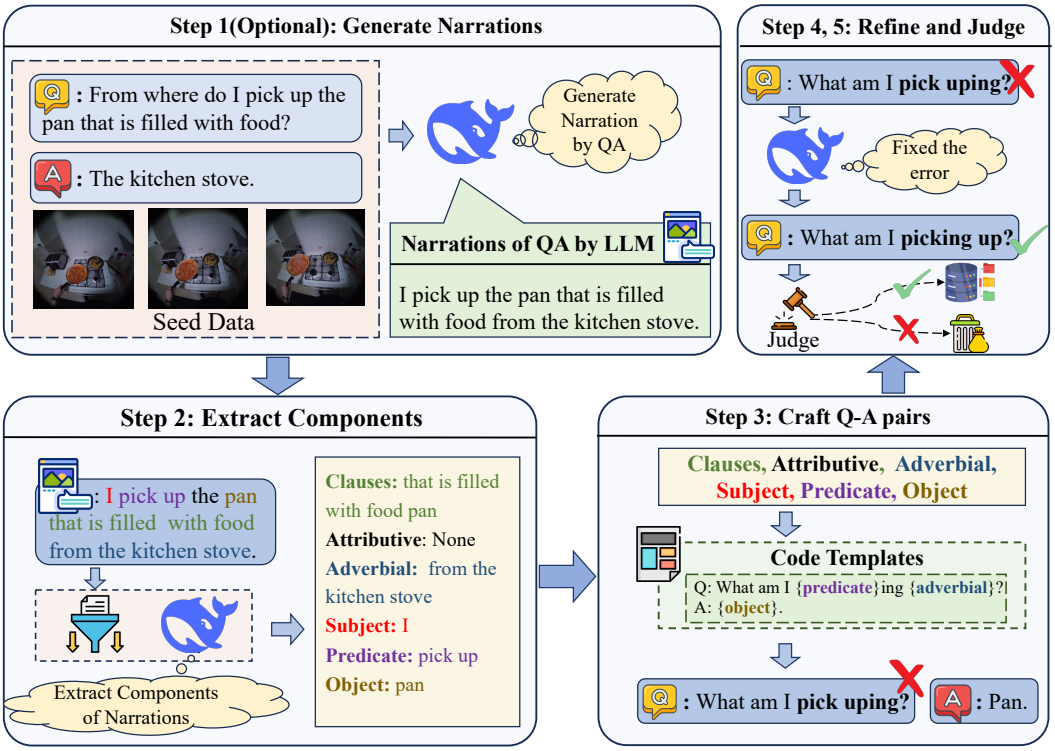

Figure 3: Overview of our egocentric QA-generation pipeline. The five image panels correspond to the five steps described in Sec. 4.2. Soft yellow rectangles (steps 1 and 4) indicate LLM-based procedures; green rectangles indicate purely code-based processes (step 3). Gradient rectangles (steps 2 and 5) denote procedures that combine LLM usage and code processing.

For videos in the HD-Epic dataset, we employ the video id and time slots provided by the vision inputs in the HD-Epic VQA benchmark directly.

4.2 QA Generation

Following the description of our method for acquiring egocentric videos and constructing time slots, we now present a detailed overview of our pipeline for generating egocentric questions and their corresponding answers. An illustrated overview of this pipeline is provided in Figure 3.

(Optional) Step 1: Generating narrations from question-answer pairs. This initial step is exclusively applied when utilizing the question-answer pairs derived from the QaEgo4D dataset (Bärmann & Waibel, 2022). For Ego4D (Grauman et al., 2022) and HD-Epic (Perrett et al., 2025), which both provide narrations directly, there is no need for Step 1. To generate narrations for subsequent steps, we prompt a capable large language model (LLM) to produce an objective, declarative sentence based on a given question-answer pair. For example, if the question is "What did I put in the refrigerator?" and the answer is "A pack of raisins," the generated sentence should be "I put a pack of raisins in the refrigerator." We observe that the employed LLM demonstrates high proficiency in performing this transliteration task, as detailed in the appendix.

Step 2: Extracting grammatical components from narrations. Using video narrations, we apply a multi-step extraction process with multiple LLM calls to identify each sentence's grammatical components. Accurate execution of this process is crucial for the subsequent question generation step, as it requires precise control over the grammatical elements. Our investigation reveals that even an advanced LLM such as Deepseek-V3.1 is unable to accurately extract all grammatical components in a single pass. To address this, we designed a chained extraction process that first extracts clauses, followed by attributives and adverbials, and finally identifies the subject, predicate, and ob-

ject. We find that the LLM can extract only a limited number of components per call when previously extracted components are removed, with detailed results provided in the appendix.

For instances where the original narration consists of multiple sentences or includes a compound sentence, we use an LLM to classify its type beforehand, retaining only the longest sentence or segment for component extraction. In cases where the narration (or preprocessed content) contains grammatical errors, these issues often become evident after the extraction process, manifesting as incomplete components—for example, the absence of a predicate.

Step 3: Crafting egocentric questions with code templates. Building on the extracted grammatical components, we proceed to construct egocentric questions following established principles. Previous studies (Mangalam et al., 2023; Cheng et al., 2024a; Zhou et al., 2025a; Chen et al., 2025) have utilized LLMs or MLLMs to generate question-answer pairs, often providing guidance and examples in their prompts to facilitate this process. These studies typically apply human-based or LLM-based post-generation filtering. Although this combination of generation and filtering can substantially improve the overall quality of questions, the outputs remain heavily influenced by the examples included in the prompts, limiting the control needed to produce truly egocentric questions. To address this, we define code templates that ensure the structural integrity of the questions, completing their content with the relevant components. We establish a minimum of five templates for each question format to enhance diversity in the generated questions.

In total, three types of questions are generated, all adhering to our principles. The first type of question includes the subject and predicate, asking about the object of the narration. The second type provides the subject, object, and its corresponding attributive, while inquiring about the predicate. The third type extends the first type by incorporating adverbials from the original sentence. Our approach maintains a high degree of control by structuring questions using predefined templates. Additionally, since grammar extraction is a process with a ground truth, our method helps to reduce errors that are difficult to detect when using MLLMs to generate questions from visual context.

Step 4: Refining questions via LLM. While combining components with code templates ensures the grammatical structure of the question, it can sometimes result in issues such as a lack of fluency or incorrect spelling. To address this, we use an LLM to refine the question while preserving its original meaning whenever possible. For example, if the generated question is "What of dough object is involved when I Add gently?", the template does not handle the attributive or capitalization properly. The LLM corrects these errors, refining the question to "What dough object is involved when I add gently?" without altering its overall meaning. We observe that the LLM can fix most issues related to fluency.

Step 5: Post-generation filtering. Previously, we implemented code to evaluate questions from existing egocentric VideoQA benchmarks. To ensure that our generated questions meet our specific requirements, we use this code to filter our own questions, removing those deemed invalid by the program. Additionally, we perform supplementary filtering steps, such as eliminating instances where the answers are contained within the questions.

4.3 Dataset analysis

EgoQuestions comprises 2,500 VideoQA instances, procured from the previously introduced video collection and QA generation processes. Our videos are sourced from two primary sources: Ego4D (Grauman et al., 2022) and HD-Epic (Perrett et al., 2025), sampled from a total of 260 raw videos. We categorize the data into three subsets based on the three types of questions previously described in Step 3 of the QA generation process. Dubbed as SPO (because we provide the Subject and Predicate in the question while inquiring about the Object), SOAP, and SPAO, the three subsets contain 1,146, 695, and 659 instances, respectively. We present the average duration of the video clips, the average word count of our questions, and other relevant information in Table 2.

5 Experiments

In this section, we evaluate both API-based and open-source vision–language models on the EgoQuestions benchmark to demonstrate the importance of using egocentric questions. We use these results to answer the following two questions:

Table 2: Dataset stats. We display the number of instances, the number of raw videos, the average clip duration, the average word count in our questions and the source of our questions for each of our subsets in this table. In the Q-A Source column, X+Y+Z denotes X questions are constructed from HD-Epic annotations, Y questions are constructed from Ego4D narrations, and Z questions are constructed from QaEgo4D Q-A pairs.

Subset	Instances	Raw Videos	Q-A Source	Avg. clip duration	Avg. Q word count
SPO	1146	207	221+738+187	3.87	5.88
SOAP	695	160	167+475+53	3.99	9.48
SPAO	659	151	152+467+40	3.51	9.88

Question 1: How well can current models answer egocentric questions about egocentric videos?

Question 2: To what extent do egocentric questions affect experimental outcomes compared with standard exocentric questions?

5.1 Evaluation

Following prior work on open-ended question answering evaluation (Cheng et al., 2024b;a), we use a capable LLM as a judge, presenting it with the question, the ground truth, and the model's response. The judge scores each answer as 0, 0.5, or 1, where 0.5 indicates a partially correct response. We provide more details, such as our evaluation prompt, in the appendix.

5.2 Models

Using EgoQuestions, we evaluate the egocentric capabilities of several recent vision–language models, including 3 API-based models and 5 open-source models. For API-based models, we evaluate GPT-5 (Seed, 2025a), doubao-seed-1.6-flash and doubao-seed-1.6(Seed, 2025b). For open-source models, we evaluate Qwen2.5-VL-7B-Instruct(Bai et al., 2025), Qwen2-VL-7B-Instruct (Wang et al., 2024), GLM-4.1V(Team et al., 2025), MiniCPM-V-4.5(Yao et al., 2025), and InternVL-3.5-8B (Wang et al., 2025), encompassing both widely recognized models such as Qwen-2.5-VL-7B-Instruct and the latest models such as MiniCPM-V-4.5 and InternVL-3.5-8B (Both of which were released in August 2025).

5.3 Generation of invalid counterparts

The question–answer pairs in EgoQuestions can effectively evaluate VLM capabilities and therefore address Question 1. However, Question 2—comparing egocentric and exocentric questions—requires further development. EgoVQA (Fan, 2019) used a similar approach to measure the performance gap between first-person and third-person questions. Since the repository linked in the original paper is no longer accessible, we examine the paper's examples and descriptions, which suggest the comparison may be unfair: first-person and third-person questions often target different aspects of the videos, making it difficult to isolate and mitigate the influence of other factors on model performance. Moreover, some Egocentric VideoQA examples in the paper contain questions that do not conform to our principles.

We argue that a rigorous comparison between egocentric and exocentric questions must meet the following requirements. First, both question types should refer to content within the same video clip. Second, paired questions for the same video should elicit identical answers and differ only in phrasing or structure. Guided by these requirements, we create exocentric question templates for comparison. Although these templates differ only slightly from the original egocentric templates, they can produce questions that provide shortcuts, allowing models to answer correctly without possessing egocentric capabilities. Apart from the template modifications, the generation process for the comparison questions is identical to that used for the original egocentric questions.

To ensure a fair comparison, we replace the code templates used in Step 3 of the QA generation pipeline with corresponding templates that produce similar but exocentric questions. This yields question pairs that share the same answer and video clip. We denote the egocentric questions as T1 (T for True) and their exocentric counterparts as F2 (F for False).

Table 3: Experimental results on the three subsets of EgoQuestions. T1 denotes the originally constructed egocentric questions, while F2 means the corresponding non-egocentric questions specifically crafted to ensure a rigorous comparison. For a subset, Δ_{acc} represents the accuracy gap between the F2 and the T1 accuracy, reflecting the impact of non-egocentric question usage.

Model	**SPO-Comparison**			**SOAP-Comparison**			**SPAO-Comparison**		
	T1	**F2**	Δ_{acc}	**T1**	**F2**	Δ_{acc}	**T1**	**F2**	Δ_{acc}
GPT-5	34.6	45.3	10.7	24.7	37.2	12.5	40.5	53.6	13.1
Doubao-1.6	32.8	45.3	12.5	24.1	34.7	10.6	36.7	46.8	10.1
Doubao-1.6 flash	30.1	41.6	11.5	19.5	27.9	8.4	32.9	44.0	11.1
Qwen2.5-VL-7B-Instruct	29.5	39.4	9.9	21.9	31.1	9.2	32.3	40.4	8.1
Qwen2-VL-7B-Instruct	30.0	38.0	8.0	18.3	27.8	9.5	34.0	38.8	4.8
InternVL3.5-8B	26.4	38.6	12.2	22.9	31.5	8.6	30.4	41.6	11.2
GLM-4.1V-Thinking	25.8	35.4	9.6	15.2	22.0	6.8	29.1	39.1	10.0
MiniCPM-V 4.5	26.1	36.4	10.3	19.8	32.4	12.6	31.6	37.8	6.2

5.4 Results

After evaluating open-source and API-based VLMs, we provide answers to the two questions that we posed above. For Question 1, current VLMs show modest performance on our open-ended egocentric questions. As shown in the T1 columns, the accuracies of VLMs range from 15% to over 40% across subsets. GPT-5 is dominant across all three subsets, outperforming the second-place model by more than 3% on the SPAO subset, while Doubao-1.6 firmly secures the position of the runner-up. Among the five open-source VLMs, no single model stands out. We can also observe the same trend across subsets for all the 8 models. All models perform best on the SPAO subset and perform the worst on the SOAP subset, proving that VLMs are still unable to grasp actions as well as they can ground objects.

For Question 2, we visualize the performance gap between ill-formatted questions and carefully crafted egocentric questions. As shown in the Δ_{acc} columns, all evaluated API-based and open-sourced VLMs exhibit a significant increase in accuracy when answering exocentric questions instead of the corresponding egocentric instances, proving that a simple shift in the format of questions can indeed result in the overrating of models' performance, leading to misjudgements concerning the egocentric abilities of a model. Take the subject SPO for example. When evaluated on the corresponding flawed questions, GLM-4.1V-Thinking, which scored lowest on the egocentric questions in SPO, scored a percentage of 35.4%, exceeding the accuracy of GPT-5 on egocentric questions. An overating percentage of more than 10% can shadow even the gap between the abilities of the model itself and GPT-5, which is among the most powerful models nowadays.

Moreover, the tendency of overrating not only appears with models that are less advanced and consist of less parameters. The performance gap appears regardless of a model's absolute performances: although GPT-5 achieves the highest accuracy on all six question formats (three egocentric and three exocentric), it still shows a Δ_{acc} of over 10% for all three subsets. The results and discussion above further demonstrate the importance of using egocentric questions in egocentric VideoQA.

6 Conclusion

In this work, we examined prior egocentric VideoQA benchmarks and identified shortcomings in their questions. We proposed three principles for egocentric questions and developed a QA-generation pipeline based on these principles to produce egocentric questions and corresponding answers. Our benchmark, EgoQuestions, comprises Q–A pairs generated by this pipeline. Experiments on EgoQuestions show that using non-egocentric questions can substantially inflate performance, demonstrating the importance of egocentric questions for evaluating models' egocentric capabilities. We will open-source our data and code to promote future research on egocentric VideoQA and related evaluations.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Leonard Bärmann and Alex Waibel. Where did i leave my keys? - episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1560–1568, June 2022.

Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. Eagle: Egocentric aggregated language-video engine. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 1682–1691, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681618. URL https://doi.org/10.1145/3664647.3681618.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2159–2167, Apr. 2025. doi: 10.1609/aaai.v39i2.32214. URL https://ojs.aaai.org/index.php/AAAI/article/view/32214.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai, 2024a. URL https://arxiv.org/abs/2410.11623.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14291–14302, June 2024b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao,

Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12943, June 2024.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

Chenyou Fan. Egovqa - an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew

Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha,

Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Woodberry. Sensecam: A retrospective memory aid. In *Proceedings of the 8th International Conference of Ubiquitous Computing (UbiComp 2006)*, pp. 177–193. Springer Verlag, September 2006. URL https://www.microsoft.com/en-us/research/publication/sensecam-a-retrospective-memory-aid/. UbiComp 2016 10-year Impact Award.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/161c94a58ca25bafcaf47893e8233deb-Paper-Datasets_and_Benchmarks.pdf.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering, 2025. URL https://arxiv.org/abs/2508.10729.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, WANG HongFa, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/31fb284a0aaaad837d2930a610cd5e50-Paper-Conference.pdf.

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

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. URL https://arxiv.org/abs/2402.13349.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 25081–25094. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4ec43957eda1126ad4887995d05fae3b-Paper-Conference.pdf.

OpenAI. Gpt-4o system card. *https://openai.com/index/gpt-4o-system-card/*, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan

Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

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

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, June 2025.

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *Int. J. Comput. Vision*, 132(11):4880–4936, May 2024. ISSN 0920-5691. doi: 10.1007/s11263-024-02095-7. URL https://doi.org/10.1007/s11263-024-02095-7.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24129–24138, June 2025.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. Egonormia: Benchmarking physical social norm understanding, 2025. URL https://arxiv.org/abs/2502.20490.

Seed. Introducing gpt-5. *https://openai.com/index/introducing-gpt-5/*, 2025a.

Seed. Seed1.6. *https://seed.bytedance.com/en/seed1*$_6$, 2025*b*.

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

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, October 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *Nat Commun 16, 5509 (2025)*, 2025.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3363–3373, June 2025a.

Wenqi Zhou, Kai Cao, Hao Zheng, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio Mayol-Cuevas, Fan Zhang, Weizhe Lin, and Junxiao Shen. X-lebench: A benchmark for extremely long egocentric video understanding, 2025b. URL https://arxiv.org/abs/2501.06835.

Chenchen Zhu, Fanyi Xiao, Andres Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20110–20120, October 2023.

A LLM USAGE

Throughout the study, we employed LLMs to assist solely in the paper writing process. The models mainly enhanced our original writing by fixing grammatical mistakes and refining the phrasing to clarify the text.

We **did not** employ LLMs to search for relevant works to put into our related works section. We looked up all these works by ourselves.

LLMs **did not** contribute to the intellectual development of the research. The ideas, analyses, and conclusions are all our own.

B IMPLEMENTATION DETAILS

We adopt a frame extracting method similar to the method employed in the VQA benchmark of HD-Epic (Perrett et al., 2025). Specifically, we process the videos at 1fps and reformat the frames to a resolution of 336*336. If the number of frames extracted exceeds 8, we uniformly sample the frames to preserve 8 frames. This method is nearly identical to the method in HD-Epic employed when assessing the model VideoLLaMA2, differing only in the upper limit of selected frames (16 for VideoLLaMA2 evaluation). We employ the same set of frames for the evaluation process of all the models, including API-based models such as GPT-5.

C DISCUSSION OF TIME SLOT SELECTION FOR EGOQUESTIONS

As mentioned in Section 4.1, previous studies (Lin et al., 2022; Plizzari et al., 2025) have utilized the Ego4D (Grauman et al., 2022) Narrations and generated time slots correspondingly. In the EgoClip pretraining dataset proposed in Lin et al. (2022), the time slot for the ith event is determined as:

$$[t_i^{start}, t_i^{end}] = [t_i - \beta_j/2\alpha, t_i + \beta_j/2\alpha], \tag{2}$$

in which t_i represents the timestamp corresponding to the ith narration. Suppose the narration is contained in the annotations of the j-th video. β_j represents the average time length between adjacent narration timestamps for this video, while α represents the average values of the β values among all the videos. We identify an α value of approximately five seconds. As the previous work claims, this format is reasonable because the length of the time slot can be controlled by both the density of identified actions on this video and the overall narration granularity. For example, if a video contains closely-distanced narrations because actions occur constantly in the video, the time slots for narrations on this video will be shorter.

However, from the equation, we expect an average length of 1 s if we generate instances for all narrations. The proposed method also always places the timestamp in the middle; this may be

effective for constructing a pretraining set, but is not required for an evaluation set. Therefore, we develop the approach described in Section 4.1. Our method significantly lengthens the time interval while still accounting for the narration density in each video.

EgoTempo (Plizzari et al., 2025) extends the time length used in EgoClip as well. Because EgoTempo requires longer videos, the authors combine up to 120 narrations to generate video clips. In contrast, our method constructs time intervals that do not overlap with any other narration, reducing contamination from surrounding narrations.

D Subset Selection for evaluation in Table 1

In Section 3.3, we evaluate only a certain subset of items from the EgoTextVQA benchmark (Zhou et al., 2025a).

We selected only a subset because the properties of the benchmark suggest a low percentage of egocentric questions. As a dataset designed to evaluate the egocentric scene-text QA ability of models, many questions ignore the recorder and focus solely on scene text. Moreover, EgoTextVQA contains over 7,000 questions; evaluating them all with our validation framework would be costly, as each Q–A instance requires multiple API calls. Therefore, we evaluate only the Gameplay subset of EgoTextVQA-Indoor. We find that more than half of the questions fail to meet our first principle, which requires egocentric questions to concern the recorder.

E Validation of LLM-based Steps

In our work, we utilize large language models (LLMs) for certain decidable tasks, such as the extraction of the grammatical components of a sentence. To ensure the trustworthiness of the outputs of these tasks, we conduct human validation of these tasks and provide the results here.

Generating Narrations from Q-A pairs. In step 1 of our QA generation pipeline, we employ an LLM to generate a narration from a question and its corresponding answer. Humans can perform this process with ease by fitting the answer in the correct place in the question. For example, we can rewrite the question "What did I put in the refrigerator?" and the answer "pack of raisins" into the objective sentence "I put a pack of raisins in the refrigerator. A correct transliteration should be fluent, preserving the information in the original Q-A pair while not introducing additional information. Upon inspecting 210 instances, we observe an accuracy of about 95.7% (201/210).

Extracting Grammatical Components. A given simple or complex sentence contains grammatical components such as a subject, a predicate, objects, attributives, adverbials, and clauses. Students who study grammar in school can often extract these components with high accuracy. However, we find that some advanced LLMs fail when prompted to extract all components in a single call. Therefore, in Step 2 of our QA-generation pipeline, we use three LLM calls, simplifying the sentence after each call. A correct extraction should place every word or phrase in its corresponding place. Upon inspecting 200 instances, we observe an accuracy of 94% (188/200) It is worth noting that the model's strictness toward attributive extraction varies over time. For example, when processing the sentence "I lift the cooking pan.", the model may determine "cooking pan" to be the object now, while determining "pan" as the object and "cooking" as an attributive related to it later. We treated both results above as correct cases during our validation.

F Prompt Hubs

F.1 Evaluation prompt

To evaluate the output of open-ended question answering, we prompt an LLM to generate a score of 0, 0.5 or 1, providing the question, ground truth and the output of the model. We employ the following prompt:

[Instruction]

- Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below.
- Your evaluation should consider correctness and helpfulness. You should especially **focus on the verbs** when evaluating the response.
- You will be given a reference answer and the assistant's answer.
- The assistant has access to an image alongwith questions but you will not be given images. Therefore, please consider only how the answer is close to the reference answer.
- Be as objective as possible. Discourage uninformative answers.
- Also, equally treat short and long answers and focus on the correctness of answers.
- After providing your explanation, you must rate the response with either 0, 0.5 or 1 by strictly following this format: `"[[rating]]"`, for example: `"Rating: [[0.5]]"`.

[Question]

{*question*}

[The Start of Reference Answer]

{*ground_truth*}

[The End of Reference Answer]

[The Start of Assistant's Answer]

{*answer*}

[The End of Assistant's Answer]

F.2 Inference Prompts

For instances constructed from the HD-Epic dataset, we utilize the following prompt wrapping: You are an expert video analyzer, and your job is to answer the open-ended question by giving only a short response. Do not give any other information.

You must give an answer, even if you are not sure.

Question:{question}

Short answer:

During InternVL3.5-8B inference, we utilize the method in the original demo code, placing "Frame-i: <image>
in front for every frame extracted.

For other VLMs, we prompt the model with the generated question directly, paired with the corresponding frames.