# EGOILLUSION: Benchmarking Hallucinations in Egocentric Video Understanding

**Anonymous ACL submission**

## Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable performance in complex multimodal tasks. While MLLMs excel at visual perception and reasoning in third-person and egocentric videos, they are prone to hallucinations, generating coherent yet inaccurate responses. We present EGOILLUSION, a first benchmark to evaluate MLLM hallucinations in egocentric videos. EGOILLUSION comprises 1,400 videos paired with 8,000 human-annotated open and closed-ended questions designed to trigger hallucinations in both visual and auditory cues in egocentric videos. Evaluations across ten MLLMs reveal significant challenges, including powerful models like GPT-4o and Gemini, achieving only 59% accuracy. EGOILLUSION lays the foundation in developing robust benchmarks to evaluate the effectiveness of MLLMs and spurs the development of better egocentric MLLMs with reduced hallucination rates. Our benchmark will be open-sourced for reproducibility[1].

## 1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have expanded their capabilities beyond image understanding to video comprehension, enabling advanced multimodal perception and reasoning (Achiam et al., 2023; Dubey et al., 2024; Ye et al., 2024b; Wang et al., 2024a; Wu et al., 2024). Depending on the camera viewpoint and observer's position, videos can be categorized as third-person (*exocentric*) videos, captured from a stationary or spectator perspective, and first-person (*egocentric*) videos, recorded from an active observer's viewpoint (Jia et al., 2024; Luo et al., 2024; Grauman et al., 2024). Egocentric videos captured from wearable devices primarily capture human-object interactions, providing rich multi-sensory information, including actions performed, object appearances, and the sounds produced during interactions (Chen et al., 2024a; Grauman et al., 2024; Kim et al., 2024; Hatano et al., 2024; Chen et al., 2024b). Unlike exocentric videos, where objects often remain static, egocentric interactions dynamically alter object states (*e.g.,* opening a bottle or turning on a device), making inference of object properties and their temporal evolution more challenging.

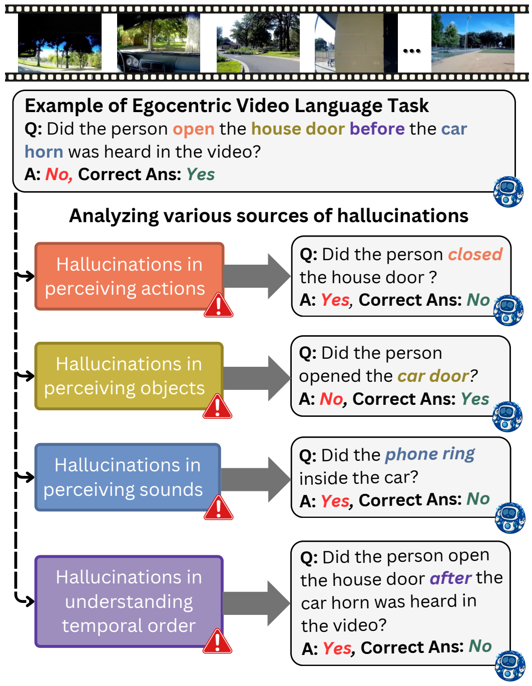


Figure 1: Illustration of various sources of hallucination encountered by MLLMs, such as Gemini (Team et al., 2024), while performing an egocentric video-language task involving temporal reasoning between two distinct events, such as a *person opening a house door* and *a car horn is heard.*

Although MLLMs demonstrate strong performance on standard image and video benchmarks (Fu et al., 2024), they remain susceptible to hallucinations, producing coherent but incorrect interpretations of sensory input that diverge from reality. As illustrated in Fig. 1, state-of-the-art MLLMs such as Gemini (Team et al., 2024) exhibit a high

[1] Please find the examples here

| Benchmark | Size | Modality | | Skills | |
| --- | --- | --- | --- | --- | --- |
| | | Vision | Audio | Perception | Reasoning |
| POPE (Li et al., 2023b) | 3k | ✓ | × | 3k ✓ | 0 × |
| HallusionBench (Guan et al., 2024) | 1.1k | ✓ | × | 0 × | 1.1k ✓ |
| MMHal-Bench (Sun et al., 2023) | 0.1k | ✓ | × | 0.05k ✓ | 0.05k ✓ |
| Bingo (Cui et al., 2023) | 0.4k | ✓ | × | 0 × | 0.4k ✓ |
| EasyDetect (Chen et al., 2024c) | 0.4k | ✓ | × | 0.4k ✓ | 0 × |
| VHTest (Huang et al., 2024) | 1.2k | ✓ | × | 0.6K ✓ | 0.6K ✓ |
| VALOR (Chen et al., 2023) | 0.2k | ✓ | × | 0.2k ✓ | 0 × |
| VideoHallucer (Wang et al., 2024b) | 1.8k | ✓ | × | 0.9k ✓ | 0.9k ✓ |
| **EGOILLUSION** *(ours)* | **8k** | ✓ | ✓ | **4.0k** ✓ | **4.0k** ✓ |

Table 1: Comparison of EGOILLUSION with existing multimodal hallucination benchmarks. EGOILLUSION covers both vision and audio modality, while having the highest number of perception and reasoning-based questions.

rate of hallucination when processing multisensory information in egocentric video, such as human actions, visual objects, and ambient sounds. Accurate perception of such elements is critical in performing common egocentric video-language tasks, including temporal reasoning between events.

**EGOILLUSION vs. Existing Benchmarks.** As shown in Table 1, we compare EGOILLUSION with existing hallucination benchmarks. Prior work has primarily focused on hallucinations in *static visual attributes* like object properties (Grauman et al., 2022; Kaul et al., 2024; Wang et al., 2023) or factual inconsistencies (Wang et al., 2024b; Guan et al., 2024), with *limited attention to video-based hallucinations*. While VideoHallucer (Wang et al., 2024b) targets *exocentric* videos, it overlooks the unique challenges of egocentric settings, such as occlusions from hand movements, action-centric narratives prone to temporal hallucinations (Grauman et al., 2022), and rich multisensory cues such as auditory cues, often misaligned by MLLMs (Su et al., 2024).

**Main Contributions.** In this work, we introduce EGOILLUSION, a benchmark designed to evaluate hallucinations in MLLMs when processing egocentric videos. EGOILLUSION includes over *1,400* egocentric videos, ranging from 30 seconds to 5 minutes, along with *8,000* human-annotated question-answer pairs. These questions assess hallucinations across diverse egocentric video-language tasks that demand advanced multimodal perception and reasoning skills. To examine hallucinations in multimodal perception, we design tasks with intricate question-answer pairs that test MLLMs' ability to infer multisensory information accurately. These tasks require models to reason about actions, sounds, and visual objects involved in human-object interactions recorded from a first-person perspective. To this end, we develop novel egocentric video-language tasks to reliably evaluate MLLMs' temporal reasoning by integrating diverse sensory cues. Additionally, we introduce hallucination questions focused on contextual and causal reasoning, which require models to infer the presence or absence of human actions, sounds, and objects before generating factually grounded responses. Our key contributions are:

- We present EGOILLUSION, the first hallucination benchmark specifically designed for egocentric video. EGOILLUSION features *8,000* question-answer pairs that capture diverse human-object interactions and enable a systematic evaluation of hallucinations across multimodal perception and understanding.
- We evaluate 10 MLLMs, including eight open-source and two proprietary models, demonstrating that state-of-the-art MLLMs exhibit a high degree of hallucinations, with the best performance of only 59% on EGOILLUSION.
- We perform extensive analysis on the models' responses and uncover key insights such as skill-wise hallucinations, challenges MLLMs face in attending multisensory input, and hallucination against diverse egocentric video-language tasks.

## 2 Related works

**Egocentric Video Understanding.** Egocentric video understanding has gained momentum with benchmarks like Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024), and EPIC-KITCHENS100 (Nasirimajd et al., 2023), which offer large-scale, annotated recordings for tasks such as activity recognition and object interaction. Multimodal datasets like QaEgo4D (Bärmann and Waibel, 2022) and EgoSchema (Mangalam et al., 2023) further enrich semantic understanding by incorporating language. Recent modeling efforts—GroundVQA (Di and Xie, 2024), Encode-Store-Retrieve (Shen et al., 2024), and R-VLM (Xu et al., 2023)—focus on long-horizon reasoning and factual consistency. However, existing benchmarks largely emphasize factual recall and recognition, lacking a systematic evaluation of hallucination. *Our work fills this gap by introducing the first benchmark designed to assess hallucination in egocentric video understanding.*

**Multimodal Large Language Models.** Recent advances in MLLMs have extended their capabilities beyond static image understanding to complex video-based perception and reasoning, incor-

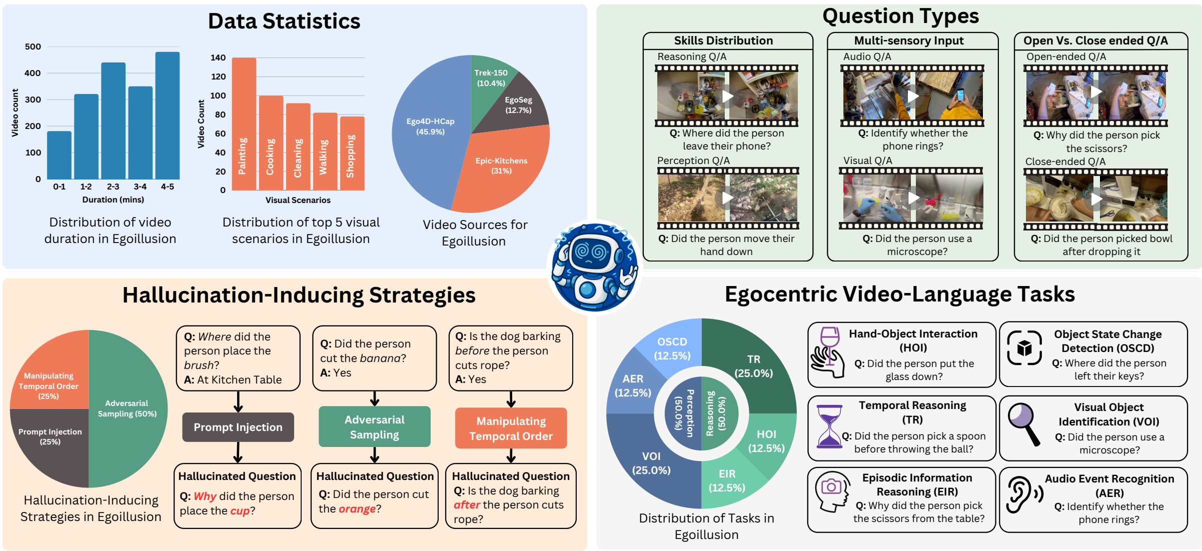


Figure 2: Overview of the EGOILLUSION benchmark. EGOILLUSION is the first hallucination benchmark for egocentric videos, featuring 8,000 human-annotated questions covering diverse egocentric video-language tasks. It presents three core challenges: (1) **Perception vs Reasoning:** distinguishing between perceptual and reasoning skills by evaluating object recognition, action understanding, and scene inference; (2) **Multisensory Inputs:** integrating visual and auditory cues, such as object appearance, human actions, and environmental sounds, to assess multimodal alignment; (3) **Question Types:** supporting both closed-ended and open-ended questions, requiring models to answer factually grounded queries while reasoning about events and interactions.

porating both visual and auditory signals (Wang et al., 2024a; Li et al., 2024b,a; Han et al., 2023). While some models rely solely on visual inputs, others explicitly integrate audio to enrich multimodal understanding (OpenBMB, 2024; Cheng et al., 2024). Most are trained primarily on third-person videos; only a few incorporate egocentric data. For instance, MiniCPM (OpenBMB, 2024) uses only third-person videos, VideoLLaMA 2 and 3 (Cheng et al., 2024; Zhang et al., 2025) mix third-person and egocentric views, and MMEgo (Ye et al., 2024a) focuses exclusively on egocentric content. Despite strong performance on standard benchmarks (Fu et al., 2024; Li et al., 2024c), *we find that these models remain susceptible to hallucinations, with the best achieving just 59% accuracy on* EGOILLUSION.

## 3 The EGOILLUSION Benchmark

### 3.1 Overview

We introduce EGOILLUSION, a novel benchmark to systematically evaluate hallucination in MLLMs across a diverse set of egocentric video-language tasks. EGOILLUSION consists of egocentric videos spanning various visual scenarios (Fig. 2), including question types requiring perceptual and reasoning skills. The benchmark features questions based on multi-sensory inputs, including visual and auditory modalities and open- and closed-ended formats. Additionally, it incorporates a range of hallucination-inducing strategies from various egocentric video-language tasks. Below, we describe the data construction pipeline of EGOILLUSION.

### 3.2 Data Collection and Filtering

We illustrate our data construction pipeline in Fig.3. The videos included in EGOILLUSION are carefully selected from a diverse collection of egocentric datasets including Ego4D-HCap (Islam et al., 2024), EgoSeg (Poleg et al., 2016), EPIC-KITCHENS (Nasirimajd et al., 2023) and Trek-150 (Dunnhofer et al., 2022), covering a wide range of visual scenarios such as meal preparation in a kitchen, painting a canvas, assembling furniture and navigating urban environments (additional details on these can be found in Appendix G). The videos in EGOILLUSION span a broad range of durations, from short clips of 30 seconds to extended recordings exceeding 5 minutes.

To ensure coverage of diverse visual content and meaningful temporal dynamics, the dataset construction of the EGOILLUSION includes a manual filtering step, which involves selecting videos that depict varied object interactions and human activities. For instance, a video showing a person transitioning from preparing ingredients to cooking and serving a meal is retained, but videos with minimal variation, such as someone stirring a pot for several minutes or walking down an empty hallway

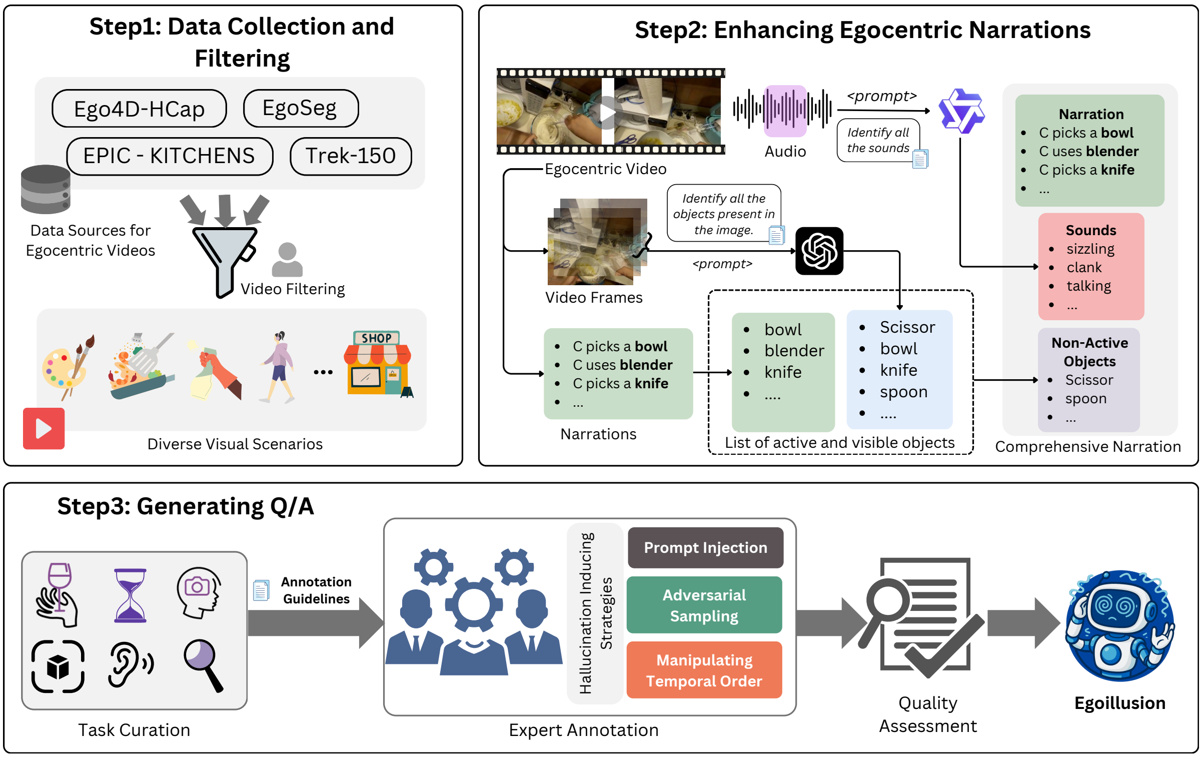


Figure 3: Illustration of the EGOILLUSION data construction pipeline. We first collect egocentric videos with detailed narrations from open-source datasets like Ego4D-HCap (Islam et al., 2024) and EPIC-KITCHENS (Nasirimajd et al., 2023), and manually filter them to ensure diverse visual scenarios (e.g., cooking, painting). We then develop an automated pipeline to enhance narrations by inferring active/inactive object states using GPT-4o (Achiam et al., 2023) and incorporating environmental sounds via Qwen2-Audio (Chu et al., 2024). Finally, we generate question-answer pairs through a rigorous human annotation process involving egocentric task design, guideline creation for inter-annotator consistency, applying hallucination-inducing strategies, and QA review.

without significant interaction, are excluded. This filtering ensures that the dataset emphasises visually and temporally rich scenarios crucial for generating complex queries and effectively evaluating hallucination in egocentric video-language models.

### 3.3 Enhancing Egocentric Narrations

While prior egocentric VQA benchmarks (Guan et al., 2024; Li et al., 2023b; Wang et al., 2024b; Chen et al., 2023) provide detailed narrations that capture a wide range of human interactions with visual elements, referred to as active objects, they often omit information about background elements, or non-active objects, that appear in the scene but are not directly interacted with. Additionally, these narrations typically lack descriptions of environmental sounds essential for comprehensive egocentric video understanding.

To address these limitations, we propose an automated pipeline to enrich egocentric narrations with visual and auditory information. As illustrated in Fig. 3, given a video $V$ with narration captions $C = \{c_1, \ldots, c_n\}$ for $n$ chronologically ordered clips, along with a global video description $D$, our method first identifies active objects, denoted by $O_I = \{o_1, \ldots, o_M\}$, based on objects the human interacts with in the narration captions. To detect non-active objects, we use GPT-4o (Achiam et al., 2023) to identify all visible objects $O_V = \{o_1, \ldots, o_P\}$ from key frames sampled from each clip. The set of non-active objects is then computed as the difference $O_S \leftarrow O_V - O_I$. In parallel, to capture environmental sounds, we use Qwen2Audio (Chu et al., 2024) to detect relevant audio cues from the soundtrack of each video clip, which results in an enriched set of egocentric narrations $C' = \{c'_1, \ldots, c'_n\}$, where each narration $c'_i$ includes not only human actions and active objects, but also associated environmental sounds and non-active objects. Finally, a manual filtering step is applied to correct potential errors and ensure the accuracy of background object and sound descriptions.

### 3.4 Generating Q/A

**Task Curation.** Leveraging insights from egocentric video corpora and our enriched narrations, we curated six egocentric video-language tasks, refined from an initial pool of 20, that target core capabilities essential for egocentric understanding, including episodic reasoning, temporal inference, and human-object interaction. Each task in EGOILLUSION is designed to assess hallucinations in either *perception* or *reasoning*, with 4,000 questions allocated to each. Perception evaluates a model's ability to interpret multi-sensory inputs by recognising human actions, sounds, and visual objects in egocentric videos. In contrast, reasoning measures the model's capacity to process this information to infer knowledge, explain causality, or make decisions (Fei et al., 2024). The selected tasks include Episodic Information Reasoning (EIR), Temporal Reasoning (TR), Human-Object Interaction (HOI), Visual Object Identification (VOI), Object State Change Detection (OSCD), and Audio Event Recognition (AER) (Additional task details are provided in Appendix D). To ensure annotation consistency and quality, we developed comprehensive, task-specific guidelines outlining objectives, expected answer formats, edge cases, and annotated examples (Additional details on the annotation guidelines are provided in Appendix E).

**Expert Annotation.** We employ expert annotators to generate question-answer pairs for each task (see Appendix E for annotator details). Annotators were provided with an annotation tool, including egocentric videos, our enriched narrations, and detailed task-specific guidelines. To create hallucinated queries, annotators were instructed to apply various hallucination-inducing strategies, such as *prompt injection, adversarial sampling, and temporal manipulation*. Detailed descriptions of these strategies are provided below (refer to Fig 2 for examples on each strategy).

*i) Prompt injection* is a simple yet effective technique for inducing hallucinations by exploiting a model's susceptibility to misleading or adversarial instructions (Liu et al., 2024). For example, given an episodic reasoning (EIR) question like "*Where did the person leave their keys?*", we inject false information by altering the question type and replacing the referenced object with one not present in the video, producing a hallucinated version such as "*Why did the person leave their hat?*" Extensive experiments reveal that MLLMs consistently fail to resist such attacks, lacking the ability to implicitly verify object presence before generating factually accurate responses.

*ii) Adversarial sampling* is employed in our benchmark to generate hallucinated queries across diverse multimodal information in egocentric videos, including human actions, sounds, and visual objects. For tasks like Hand-Object Interaction (HOI), we create hallucinated counterparts by replacing the active object (*i.e.,* the one being interacted with) with a non-active object in the scene. Using this strategy, we ensure that the hallucinated action-object pairs are scene-aware, making them harder to defend against.

*iii) Manipulating temporal order* is used in our benchmark to generate hallucinated queries by altering the sequence of events defined by human-object interactions in egocentric videos. By reordering these interactions, we create mismatches between actions and the corresponding sounds they produce. This results in temporally inconsistent yet scene-plausible queries, increasing the difficulty for models in detecting hallucinations.

**Quality Assessment.** To ensure the quality and consistency of the annotations, we conducted a structured quality assessment protocol involving iterative feedback and reliability checks. After initial annotation, all question-answer (QA) pairs were reviewed through a back-and-forth process between expert annotators and authors. Annotators were encouraged to flag ambiguous cases or annotation uncertainties, which were then discussed in weekly review meetings. To quantitatively assess annotation reliability, we randomly selected 1,000 QA pairs across all six tasks and had them cross-verified by expert reviewers. We measured inter-annotator agreement using Krippendorff's Alpha, a standard metric for multi-rater agreement in benchmark construction (Thrush et al., 2022; Li et al., 2023a), and observed an average alpha score of *0.78*, indicating substantial agreement across perception and reasoning tasks.

## 4 Experimental Setup

We first describe the baselines used to evaluate hallucination performance and then outline the human evaluation setup.

**Baselines.** We benchmark a range of MLLMs, including eight open-weight and closed-source models, such as Gemini-1.5 (Team et al., 2024) and GPT-4o (Achiam et al., 2023). These models are selected to cover a wide variety of factors, including *model size* (LLaVa-OV (Li et al., 2024a) contains 0.5B parameters, whereas VideoLLaMA2 (Cheng et al., 2024) consists of 7B

| Models | Size | Ego | Modality | | Reasoning Skills | | | Perception Skills | | | Avg (↑) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Vision | Audio | EIR (↑) | TR (↑) | HOI (↑) | VOI (↑) | OSCD (↑) | AER (↑) | |
| *Human Evaluation* | | | | | | | | | | | |
| Human | | | | | 80.1$_{\pm0.2}$ | 86.5$_{\pm0.2}$ | 84.2$_{\pm0.4}$ | 88.4$_{\pm0.5}$ | 91.1$_{\pm0.3}$ | 86.3$_{\pm0.2}$ | 86.1$_{\pm0.3}$ |
| *Open-Source Models* | | | | | | | | | | | |
| Qwen2.5VL (Bai et al., 2025) | 3B | × | ✓ | × | 50.1$_{\pm0.3}$ | <u>67.3</u>$_{\pm0.2}$ | 54.6$_{\pm0.4}$ | 56.3$_{\pm0.1}$ | 51.1$_{\pm0.3}$ | - | 55.8$_{\pm0.2}$ |
| VideoLlama3 (Zhang et al., 2025) | 8B | ✓ | ✓ | × | 52.1$_{\pm0.4}$ | 59.9$_{\pm0.3}$ | 62.7$_{\pm0.2}$ | 63.9$_{\pm0.5}$ | 53.2$_{\pm0.1}$ | - | 58.3$_{\pm0.3}$ |
| InternVideo (Wang et al., 2025) | 8B | ✓ | ✓ | × | 51.4$_{\pm0.4}$ | 64.3$_{\pm0.1}$ | <u>65.5</u>$_{\pm0.2}$ | 60.8$_{\pm0.3}$ | 51.7$_{\pm0.2}$ | - | 58.7$_{\pm0.3}$ |
| LLaVa-NEXT (Li et al., 2024b) | 7B | × | ✓ | × | 50.1$_{\pm0.2}$ | 58.4$_{\pm0.5}$ | 64.1$_{\pm0.1}$ | 56.8$_{\pm0.3}$ | **61.9**$_{\pm0.4}$ | - | 58.2$_{\pm0.2}$ |
| LLaVa-OV 0.5B (Li et al., 2024a) | 0.5B | ✓ | ✓ | × | 51.2$_{\pm0.3}$ | 64.5$_{\pm0.1}$ | 61.8$_{\pm0.4}$ | 60.5$_{\pm0.2}$ | 52.4$_{\pm0.5}$ | - | 58.1$_{\pm0.3}$ |
| LLaVa-OV (Li et al., 2024a) | 7B | ✓ | ✓ | × | 51.2$_{\pm0.4}$ | **67.5**$_{\pm0.2}$ | 62.9$_{\pm0.3}$ | 58.5$_{\pm0.1}$ | 50.3$_{\pm0.5}$ | - | 58.1$_{\pm0.2}$ |
| ImageBind-LLM (Han et al., 2023) | 7B | × | ✓ | ✓ | 55.2$_{\pm0.3}$ | 65.6$_{\pm0.4}$ | 61.6$_{\pm0.2}$ | 52.9$_{\pm0.1}$ | 51.6$_{\pm0.3}$ | 52.2$_{\pm0.5}$ | 57.3$_{\pm0.2}$ |
| MiniCPM (OpenBMB, 2024) | 8B | × | ✓ | ✓ | **57.3**$_{\pm0.4}$ | 47.3$_{\pm0.1}$ | **66.9**$_{\pm0.5}$ | 69.5$_{\pm0.3}$ | <u>58.4</u>$_{\pm0.2}$ | 50.1$_{\pm0.4}$ | <u>58.9</u>$_{\pm0.3}$ |
| VideoLlama2 (Cheng et al., 2024) | 7B | ✓ | ✓ | ✓ | <u>56.1</u>$_{\pm0.3}$ | 38.9$_{\pm0.2}$ | 40.2$_{\pm0.5}$ | 41.2$_{\pm0.4}$ | 56.8$_{\pm0.1}$ | **52.6**$_{\pm0.3}$ | 47.6$_{\pm0.2}$ |
| *Closed-Source Models* | | | | | | | | | | | |
| Gemini-Pro (Team et al., 2024) | - | - | ✓ | ✓ | 51.4$_{\pm0.2}$ | 60.8$_{\pm0.3}$ | 61.8$_{\pm0.5}$ | <u>68.1</u>$_{\pm0.4}$ | 56.5$_{\pm0.1}$ | <u>52.5</u>$_{\pm0.3}$ | **59.4**$_{\pm0.2}$ |
| GPT-4o (Achiam et al., 2023) | - | - | ✓ | × | 53.2$_{\pm0.3}$ | 47.5$_{\pm0.2}$ | 66.7$_{\pm0.4}$ | **73.9**$_{\pm0.5}$ | 58.4$_{\pm0.1}$ | - | 58.8$_{\pm0.3}$ |

Table 2: Performance comparison of various MLLMs on EGOILLUSION across egocentric video-language tasks: Episodic Information Reasoning (**EIR**), Temporal Reasoning (**TR**), Human-Object Interaction (**HOI**), Visual Object Identification (**VOI**), Object State Change Detection (**OSCD**), and Audio Event Recognition (**AER**). We indicate whether the models were trained on egocentric video data and whether they leverage both vision and audio modalities. The best-performing models for each task are highlighted in **bold**, while the second-best scores are <u>underlined</u>.

parameters). They also vary in the *video type* used during training (ImageBind-LLM (Han et al., 2023) is trained solely on exocentric videos, while VideoLLaMA3 (Zhang et al., 2025) and InternVideo (Wang et al., 2025) are jointly trained on both exocentric and egocentric videos). Finally, the models differ in their *multisensory input capabilities* — LLaVa-Next (Li et al., 2024b) and LLaVa-OV (Li et al., 2024a) process videos without audio, in contrast to models like Gemini-1.5 (Team et al., 2024), which process both video and audio signals (see Appendix B for additional details).

**Hallucination Evaluation.** We conduct separate evaluations for both close-ended and open-ended questions. For close-ended questions, which require binary yes/no answers, we follow prior video hallucination benchmarks such as VideoHallucer (Wang et al., 2024b) by applying string matching to convert model responses into either "*Yes*" or "*No*." For open-ended questions, we adopt a two-step approach: first, we determine whether the model implicitly assumes the presence of an object using an LLM-as-judge framework (Zheng et al., 2023) with GPT-4o (Achiam et al., 2023) (to reduce model bias, we also use Gemini-Pro for LLM-as-judge); second, we independently assess the factual correctness of the response. Consistent with previous hallucination benchmarks, we report accuracy as the primary metric, where lower accuracy indicates a higher degree of hallucinations.

**Human Evaluation.** We recruited three English-proficient individuals to evaluate our benchmark, where each individual had strong foundational knowledge of computer vision. To reduce potential evaluator bias, we randomized the order of the question-answer pairs, ensuring that correct and hallucinated responses did not appear consecutively. Inter-annotator reliability was measured using the Pearson correlation coefficient, yielding a moderate agreement score of *0.58*.

## 5 Results

### 5.1 Main Results

We benchmark ten state-of-the-art MLLMs on EGOILLUSION and present the results in Table 2. Below, we summarize the key findings:

*i)* EGOILLUSION *presents a significant challenge*, exposing the vulnerability of current MLLMs to hallucination. We find that existing models struggle to defend against hallucinations induced by EGOILLUSION. For instance, the best-performing model, Gemini-Pro, achieves 59.4% accuracy, while human performance on the benchmark is 86.1%, revealing a gap of *26.7%*.

*ii) Minimal performance gap between open- and closed-weight model.* Unlike other benchmarks, EGOILLUSION reveals only a small performance gap between open- and closed-weight models. In Table 2, we show that the best open-weight model, VideoLlama3, achieves an accuracy of 58.3%, while the best closed-weight model, Gemini-Pro, reaches 59.4%, a marginal difference of *1%*.

| Models | PI (↑) | AS (↑) | MTO (↑) |
|---|---|---|---|
| | *Open-Weight Models* | | |
| ImageBind-LLM | $54.5_{\pm0.3}$ | $61.6_{\pm0.4}$ | $65.6_{\pm0.2}$ |
| Qwen2.5VL | $53.2_{\pm0.2}$ | $52.8_{\pm0.3}$ | $67.3_{\pm0.5}$ |
| VideoLlama3 | $60.1_{\pm0.4}$ | $66.0_{\pm0.2}$ | $59.9_{\pm0.3}$ |
| LLaVa-NEXT | $58.0_{\pm0.1}$ | $65.3_{\pm0.5}$ | $58.4_{\pm0.3}$ |
| LLaVa-OV 0.5B | $56.5_{\pm0.3}$ | $57.2_{\pm0.4}$ | $64.5_{\pm0.2}$ |
| LLaVa-OV | $54.8_{\pm0.2}$ | $56.8_{\pm0.3}$ | $67.5_{\pm0.4}$ |
| MiniCPMo-2.6 | $58.4_{\pm0.5}$ | $51.0_{\pm0.2}$ | $47.3_{\pm0.3}$ |
| VideoLlama2 | $58.9_{\pm0.3}$ | $51.0_{\pm0.4}$ | $38.9_{\pm0.2}$ |
| | *Closed-Source Models* | | |
| Gemini-Pro | $53.9_{\pm0.4}$ | $64.9_{\pm0.2}$ | $60.8_{\pm0.5}$ |
| GPT-4o | $54.2_{\pm0.3}$ | $62.1_{\pm0.1}$ | $59.7_{\pm0.3}$ |

Table 3: Performance comparison of various MLLMs across diverse hallucination-inducing strategies employed in EGOILLUSION, including prompt injection (PI), Adversarial Sampling (AS), and Manipulating Temporal Order (MTO).

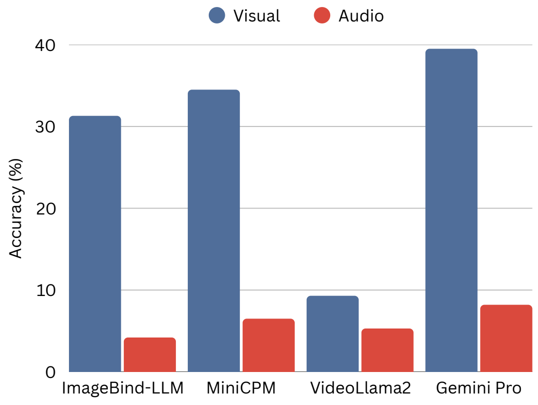


Figure 4: Performance comparison on confounding pairs generated from the videos Q/A sourced from EGOILLUSION across visual and audio modality.

*iii) Minimal performance gap between small and large MLLMs.* Unlike conventional benchmarks where larger models typically outperform smaller ones, EGOILLUSION reveals that model size alone does not consistently mitigate hallucinations, *e.g.,* the small LLaVA-OV 0.5B model achieves 58.1% average accuracy, matching the performance of its larger counterpart, LLaVA-OV 7B, suggesting that the hallucinations introduced by EGOILLUSION are not easily mitigated by scaling model size.

*iv) MLLMs hallucinate less on perception-based tasks than on reasoning tasks.* As shown in Table 2, MLLMs hallucinate less on perception-based tasks (Visual Object Identification (VOI) and Audio Event Recognition (AER)) compared to reasoning tasks (Temporal Reasoning (TR) and Episodic Information Reasoning (EIR)). For example, the best-performing model, Gemini-Pro, achieves 68.1% accuracy on VOI and 58.3% on AER, but only 60.8% on TR and 51.4% on EIR—*a gap of over 7%*. This suggests that hallucinations are more prevalent when models are asked to perform complex reasoning rather than perception.

### 5.2 Ablation On Hallucination Inducing Strategies

Building on these findings, we further examine how different hallucination-inducing strategies affect MLLM performance on EGOILLUSION. Table 3 compares the performance of various MLLMs under different hallucination-inducing strategies employed in the EGOILLUSION. Overall, models tend to perform close to random guess across all strategies, highlighting their consistent vulnerability to hallucinations in egocentric video understanding. Among open-weight models, MiniCPM and VideoLlama2 perform the worst, particularly under the Manipulating Temporal Order (MTO) strategy, where their scores drop to 47.3% and 38.9%, respectively, indicating significant difficulty in understanding chronological ordering in unique egocentric events. For closed-weight models, Gemini-Pro and GPT-4o perform reasonably well compared to open-weight models but remain susceptible to hallucinations induced by Prompt Injection (PI), where they achieve the lowest score (53.9%), indicating that these MLLMs are vulnerable to misleading prompts, likely due to learned biases from pretraining data that make them more susceptible to hallucinated inputs.

### 5.3 Which modality does MLLMs attend to?

Motivated by the near-random performance of current MLLMs on our benchmark, we further investigate which modality (audio or visual) these models primarily attend to while understanding egocentric videos. We conduct an experiment by randomly selecting 200 video clips from EGOILLUSION and generating confounding pairs to isolate the contribution of each modality. For the audio modality, we synthetically add unrelated background sounds; for the visual modality, we replace the main object in the query with a random object. A model's response is considered correct only if it answers both versions of the confounding pair correctly. As shown in Fig. 4, when evaluated on MLLMs that process both modalities, we find a significant drop in performance below 50%, on both types of pertur-

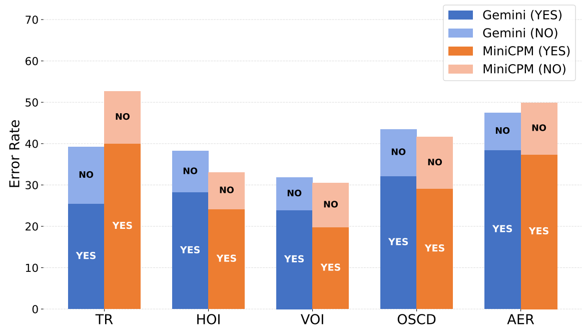


Figure 5: Distribution of "*Yes*" and "*No*" responses of Gemini-1.5 Pro and MiniCPM in the hallucinated responses for closed-ended questions. We observe that the model is inclined towards affirmative responses in hallucinated outputs.

bations. Notably, the *performance degradation is more severe for audio*, with a 32% drop for Gemini and 28% for MiniCPM. These results demonstrate that while MLLMs struggle with both modalities, they especially fail to leverage audio cues, instead relying heavily on language priors, leading to hallucinated responses.

## 5.4 Error Analysis

Next, we conduct a detailed error analysis with a focus on response biases and the failure cases.

**Yes/No Bias.** Fig. 5 presents a quantitative analysis of how often Gemini-1.5 Pro and MiniCPM respond with "*Yes*" or "*No*" when generating hallucinated responses in closed-ended tasks within EGOILLUSION. We observe that despite differing hallucination rates, both models exhibit a significantly higher proportion of "*Yes*" responses compared to "*No*" across various tasks, *e.g.*, in egocentric video-language tasks such as Visual-Object Identification (VOI), where both models show similar hallucination rates, we find that they still demonstrate a strong bias toward "*Yes*" responses. A similar pattern emerges in Temporal Reasoning (TR), where the models differ in their hallucination rates but still predominantly produce "*Yes*" responses. This trend remains consistent across other tasks, as shown in Fig. 5, indicating the models' inclination toward affirmative responses in hallucinated outputs.

**Finegrained Error Analysis.** We conduct a manual error analysis on 1,000 incorrect responses, representing 12.5% of the total benchmark samples, uniformly sampled across all six tasks in EGOILLUSION. Fig. 6 presents a detailed breakdown of the different types of errors observed in responses generated by Gemini 1.5 Pro (Team et al., 2024) and

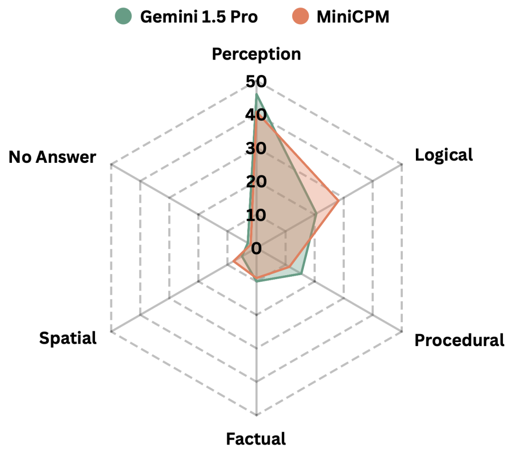


Figure 6: An illustration of the different types of errors observed in incorrect responses from Gemini 1.5 Pro (Team et al., 2024) and MiniCPM (OpenBMB, 2024). Additional details on the various error types can be found in Appendix J.

MiniCPM (OpenBMB, 2024) on EGOILLUSION. The primary source of errors for both models is *perception*, accounting for 48.6% of Gemini 1.5 Pro's mistakes and 43.7% of MiniCPM's. This is largely driven by hallucination-inducing questions in EGOILLUSION, revealing the models' difficulty in accurately perceiving entities in the video before generating factually grounded responses. In addition, *logical* and *procedural* errors make up a substantial share of the failures, indicating that even when models identify relevant entities correctly, they often fall short in applying the complex reasoning needed for accurate answers. Overall, this analysis underscores the critical need for improved perceptual understanding in egocentric video tasks.

# 6 Conclusion

In this paper, we introduced EGOILLUSION, the first comprehensive benchmark specifically designed to evaluate hallucination in MLLMs within egocentric video understanding. Our benchmark features over 1,400 egocentric videos and 8,000 carefully annotated question-answer pairs designed to systematically trigger and assess hallucinations across diverse scenarios involving audio and visual perception and complex reasoning. Experimental results across ten SOTA MLLMs reveal significant vulnerabilities, demonstrating that current models, regardless of scale or training modality, are highly susceptible to hallucinations, achieving accuracies close to random guessing. By introducing novel hallucination inducing techniques, EGOILLUSION provides insights into the MLLM's limitations and offers a roadmap for future research.

## 7 Limitation and Future Work

In this section, we highlight a few limitations and future directions:

- Our benchmark, EGOILLUSION reveals that existing Multimodal Large Language Models (MLLMs) exhibit a high rate of hallucination when evaluated on egocentric video understanding tasks. In future work, we plan to develop robust hallucination mitigation strategies tailored specifically for this domain.
- While the current version of our benchmark evaluates model performance on visual and non-speech auditory cues (e.g., background sounds) in egocentric videos, it does not yet cover speech signals. As egocentric videos often contain conversations, we aim to extend our benchmark to include the speech modality in future iterations, enabling more comprehensive evaluations and analysis.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv*.

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. 2025. Qwen2.5-vl technical report.

Leonard Bärmann and Alex Waibel. 2022. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *CVPR*.

Changan Chen, Kumar Ashutosh, Rohit Girdhar, David Harwath, and Kristen Grauman. 2024a. Soundingactions: Learning how actions sound from narrated egocentric videos.

Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. 2024b. Action2sound: Ambient-aware generation of action sounds from egocentric videos.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024c. Unified hallucination detection for multimodal large language models. *arXiv*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv*.

Shangzhe Di and Weidi Xie. 2024. Grounded question-answering in long egocentric videos. In *CVPR*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv*.

Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. 2022. Visual object tracking in first person vision. *IJCV*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv*.

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

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Imagebind-llm: Multi-modality instruction tuning.

Masashi Hatano, Ryo Hachiuma, Ryo Fujii, and Hideo Saito. 2024. Multimodal cross-domain few-shot learning for egocentric action recognition.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multi-modal large language models. *arXiv*.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *CVPR*.

Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M. Rehg, Vamsi Krishna Ithapu, and Ruohan Gao. 2024. The audio-visual conversational graph: From an egocentric-exocentric perspective. In *CVPR*.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *CVPR*.

Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. 2024. Palm: Predicting actions through language models.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark.

Xinyu Li et al. 2023a. Vl-coco: A benchmark for visio-linguistic compositional reasoning. In *ICCV*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024. Prompt injection attack against llm-integrated applications.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. 2024. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*.

Amirshayan Nasirimajd, Simone Alberto Peirone, Chiara Plizzari, and Barbara Caputo. 2023. Epic-kitchens-100 unsupervised domain adaptation challenge: Mixed sequences prediction. *arXiv*.

OpenBMB. 2024. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone. Accessed: 2025-03-07.

Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact cnn for indexing egocentric videos. In *WACV*.

Junxiao Shen, John J Dudley, and Per Ola Kristensson. 2024. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *IEEE ISMAR*.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2024. From vision to audio and beyond: a unified model for audio-visual representation and generation. In *ICML*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*.

Tristan Thrush et al. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv*.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. 2023. Retrieval-based video language model for efficient long video question answering. *arXiv*.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. 2024a. Mm-ego: Towards building egocentric multimodal llms. *arXiv*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*.

## A Appendix

In the Appendix, we provide:

1. Section B: Baseline Details
2. Section C: Other Benchmark Details
3. Section D: Tasks
4. Section E: Annotator Details
5. Section F: Annotation Guidelines
6. Section G: Data Source and Filtering

## B Baseline Details

**ImageBind-LLM**[2] (Han et al., 2023) ImageBind-LLM is built on a 7B-parameter LLaMA base, augmented with a learnable bind network to align ImageBind's image encoder with LLaMA. It is trained solely on exocentric image-text pairs. Although its training data contains only images (without audio), its unified embedding space allows it to handle audio, video, and 3D point cloud inputs during inference.

**VideoLlama2**[3] (Cheng et al., 2024) VideoLlama2 is a video-language model with 7B parameters that leverages a LLaMA-based language model. It processes video inputs comprising both visual frames and audio. The model is trained on large-scale exocentric video–text datasets, where the video data is provided with audio.

**MiniCPM**[4] (OpenBMB, 2024) MiniCPM is a multimodal large language model with 8B parameters. It accepts live video frames along with synchronized speech inputs, making it great for real-time multimodal live streaming, especially on edge and mobile devices. The model is trained on a variety of datasets that include exocentric video data. The training data comprises video sequences with audio, which allows for effective vision–speech alignment and a richer multimodal understanding.

**InternVideo**[5] (Wang et al., 2025) InternVideo2.5 is built on a 7B-parameter base using InternLM2.5-7B as its language adapter. It takes video inputs - sequences of video frames accompanied by text instructions, with a focus on visual content (without audio). The model is trained on a variety of egocentric and exocentric video datasets, covering both short and long video contexts.

**Qwen2.5VL**[6] (Bai et al., 2025) Qwen2.5VL is a multi-modal model with roughly 3B parameters that uses a Qwen-based language model as its adapter. It processes inputs from video, where the data includes visual frames, allowing multi-modal comprehension. The model is pre-trained on large-scale exocentric video-text datasets. It's training setup ensures that Qwen2.5VL is good at interpreting visual information for tasks like video captioning and question answering.

**VideoLlama3**[7] (Zhang et al., 2025) VideoLlama3 is an advanced video-language model built with 8B parameters, using a LLaMA-based language model as its foundation. It accepts video inputs that includes visual frames, which allows it to capture temporal cues. The model is trained on extensive egocentric and exocentric video datasets. It's training methodology allows VideoLlama3 to perform very well at real-time video understanding and multi-modal reasoning tasks.

**LLaVa-NEXT**[8] (Li et al., 2024b) LLaVa-NEXT is a vision-language model having 7B parameters and is built on a LLaMA-derived language model adapter. It accepts video inputs as image frames and text queries, focusing exclusively on visual content without audio. The model is trained on large-scale exocentric image-text datasets. Its training data comprises high-quality images, which ensures accurate visual-text alignment and great performance on tasks such as image captioning and visual question answering.

**LLaVa-OneVision**[9] (Li et al., 2024a) LLaVa-OneVision is a vision-language model having approximately 7B parameters and is built on a LLaMA-based language model. It takes static image inputs along with text for rich visual-text interactions. The model is trained on egocentric and exocentric image-text datasets. Its training data consists of images paired with text, enabling it to deliver high performance on tasks like image captioning, retrieval, and dialogue generation. We have tested our benchmark on both 0.5B and 7B parameter versions of LLaVa-OneVision.

**Gemini-1.5-Pro** (Team et al., 2024) Gemini 1.5

[2] https://github.com/dynamic-superb/multimodal-llama
[3] https://github.com/DAMO-NLP-SG/VideoLLaMA2
[4] https://github.com/OpenBMB/MiniCPM
[5] https://github.com/OpenGVLab/InternVideo
[6] https://github.com/QwenLM/Qwen2.5-VL
[7] https://github.com/DAMO-NLP-SG/VideoLLaMA3
[8] https://github.com/LLaVA-VL/LLaVA-NeXT
[9] https://github.com/LLaVA-VL/LLaVA-NeXT/blob/main/docs/LLaVA_OneVision_Chat.md

Pro is a proprietary multimodal model by Google. It is state-of-the-art on many video benchmarks. It is capable of processing and reasoning over extremely long contexts, up to 10 million tokens. It outperforms its competitors in long-document QA, video and audio analysis, and retrieval tasks.

**GPT-4o** (Achiam et al., 2023) GPT-4o is OpenAI's latest multimodal model capable of processing text, images, and audio natively, offering faster and more accurate responses across modalities. Compared to previous versions, GPT-4o demonstrates improved reasoning abilities, enhanced real-time interaction, and better alignment with user intent, making it particularly suitable for interactive and perception-heavy tasks.

## C Other Benchmark Details

**POPE** (Li et al., 2023b) POPE is an image-based hallucination evaluation dataset consisting of 3000 questions over 500 images. It is designed to assess object hallucinations using a binary QA format, focusing on detecting whether a specified object is present or hallucinated. The dataset is constructed from exocentric image data and does not incorporate adversarial testing.

**HallusionBench** (Guan et al., 2024) HallusionBench supports both image and video modalities and comprises 1129 questions over 346 instances. It evaluates multiple hallucination aspects, such as object, relational, and semantic errors, using an LLM-based evaluation protocol. The data is exocentric, and the benchmark does not include adversarial components.

**MMHal-Bench** (Sun et al., 2023) MMHal-Bench is an image-based evaluation benchmark with 96 questions on 96 images. It focuses on hallucinations in object, relational, and semantic details, employing an LLM-based evaluation approach. The dataset uses exocentric imagery and does not involve adversarial testing.

**Bingo** (Cui et al., 2023) Bingo is an image-focused benchmark featuring 370 questions across 370 images. It assesses hallucination issues, particularly object-level and semantic inconsistencies, using an LLM-based evaluation method combined with an adversarial component, making it more challenging to detect hallucinations reliably.

**EasyDetect** (Chen et al., 2024c) EasyDetect is an image-based hallucination detection dataset with 420 questions over 420 images. It targets object, relational, and semantic hallucinations using an LLM-based evaluation framework. The data is exocentric, and the benchmark does not include adversarial settings.

| TASK | # QUES | TYPE |
|---|---|---|
| Episodic Information Reasoning | 1000 | Open-ended |
| Temporal Reasoning | 2000 | Closed-ended |
| Hand-Object Interaction | 1000 | Closed-ended |
| Visual Object Identification | 2000 | Closed-ended |
| Episodic Information Extraction | 1000 | Closed-ended |
| Audio Event Recognition | 1000 | Closed-ended |

Table 4: Distribution of number of questions and their type for each task

**VHTest** (Huang et al., 2024) VHTest is an image dataset containing 1200 questions on 1200 images, designed to evaluate hallucinations in visual outputs. It focuses on assessing object and semantic hallucination types through an LLM-based evaluation method without adversarial enhancements. The images are exocentric in nature.

**VALOR** (Chen et al., 2023) In the hallucination evaluation context, VALOR is an image-based dataset with 211 questions on 211 images. It is used to measure object, relational, and semantic hallucinations via an LLM-based evaluation protocol, relying on exocentric imagery and without adversarial testing.

**VideoHallucer** (Wang et al., 2024b) VideoHallucer is a video-based benchmark with 1800 questions across 948 videos. It comprehensively covers a wide range of hallucination types, including object-relation, semantic, temporal, extrinsic factual, and non-factual hallucinations. The evaluation is performed using a binary QA method with an adversarial component, ensuring robust assessment of LVLMs' performance on dynamic video content.

## D Tasks

**Episodic Information Reasoning (EIR)** evaluates MLLMs' ability to accurately track objects and their interactions over time in egocentric videos and furter reason over this information. This task is particularly challenging in egocentric settings, where the first-person perspective creates a dynamic field of view with objects frequently entering, exiting, and being manipulated through a series of actions. In this task, models must answer "how,", "what", "why,", "where" (not exclusive to these types) questions about objects that appeared in the video while correctly identifying when questions refer to objects that were never present. The task specifically targets hallucination tendencies by including plausible but non-existent objects that fit the scene context, testing whether models can resist generating false information about actions that never occurred.

**Examples:**

- Why did the person push the bicycle?
- Where did the person place the pliers?
- What did the person do with their hand?

The answers to these are open-ended but grounded in the visual and acoustic environment of the agent.

**Temporal Reasoning (TR)** evaluates MLLMs' ability to track chronological relationships between events in egocentric videos. This task tests whether models can accurately determine the temporal order of actions that are separated by several intervening events, challenging them to maintain a coherent understanding of the activity timeline. In egocentric settings, where the first-person perspective creates a continuous stream of interactions, properly sequencing events becomes particularly challenging as objects and actions flow in and out of view. The task presents questions using "before/after" temporal operators to probe if models can correctly identify the relative ordering of events without hallucinating plausible but incorrect sequences.

**Examples:**

- Did the person open the gate after passing the broom from his right hand to the left hand?
- Did the person wash the car after putting the hose down?

The answers to these are closed-ended and can be either Yes or No

**Hand-Object Interaction (HOI)** evaluates MLLMs' ability to detect physical actions in egocentric videos. This task challenges models to distinguish between actual hand-object interactions that occurred in the video and visually similar but non-occurring actions. By presenting pairs of original actions (e.g., "picking up an object") alongside contrastive alternatives (e.g., "throwing an object"), the task tests whether models hallucinate plausible interactions or accurately recall the specific physical actions that were performed.

**Examples:**

- Did the person pick a cooking spoon?
- Did the person carry the timber?

The answers to these are closed-ended and can be either Yes or No

**Object State Change Detection (OSCD)** evaluates MLLMs' ability to reason about state changes and action completeness in egocentric videos through yes/no questions. Unlike Episodic Information Reasoning, which tests open-ended reasoning through "how," "why," and "where" questions, this task uses binary questions to assess whether models can accurately track object state transformations and recall this information when requested. The task challenges models to identify complete action pairs (where objects return to their initial state, like opening and closing a fridge) versus incomplete actions (where state changes remain unresolved, such as removing an item without replacing it).

**Examples:**

- Did the person insert the screw after picking it up?
- Did the person put down the blender jar after taking it?

The answers to these are closed-ended and can be either Yes or No

**Visual Object Identification (VOI)** evaluates MLLMs' ability to correctly determine which objects were involved in specific activities within egocentric videos. This task challenges models to distinguish between objects that were genuinely part of an activity (e.g., eggs used while cooking) and plausible but absent objects (e.g., carrots that would fit the cooking scenario but never appeared). By providing an activity context through visual captions, the task creates a particularly challenging scenario for hallucination detection, as models must resist the temptation to associate semantically related but absent objects with the identified activity.

**Examples:**

- Did the person remove the plug from the fuel pipe?
- Did the person peel the potato with a knife?

The answers to these are closed-ended and can be either Yes or No

**Audio Event Recognition (AER)** evaluates MLLMs' ability to distinguish between actual audio cues and plausible but non-existent background sounds in egocentric videos. This task challenges models to identify appropriate moments where synthetic background sounds could be added that are coherent with the visual scene but not inherently produced by the actions being performed. By requiring models to determine which background sounds would be plausible in specific contexts (e.g., a phone ringing during cooking or distant dog barking when near a window), the task tests whether models can accurately separate observed audio information from inferred possibilities. This is particularly revealing in egocentric videos, where the first-person perspective often includes rich environmental audio that models may hallucinate based on visual cues alone.
**Examples:**

- Did you hear the sound of birds chirping
- Did you hear the sound of the cash register?

The answers to these are closed-ended and can be either Yes or No

# E Annotator Details

We employed five experts to annotate the data, which included 3 males and 2 females. The experts are MS/PhD students witha strong foundational understanding of computer vision. All annotators had prior experience with video annotation tasks and were familiar with the challenges of egocentric vision.

Before beginning the annotation process, annotators were given training sessions to ensure consistency in their understanding of hallucination categories and annotation guidelines. This training included an overview of hallucination categories, followed by short exercises in which they were asked to annotate some examples, which were reviewed and discussed.

The annotation process was conducted over a period of 4 weeks, with regular meetings to address doubts and calibrate their understanding of the hallucination categories . Annotators were compensated fairly for their expertise and time commitment. For conducting annotations, we got the approval from our Institution Review Board (IRB)

# F Annotation Guidelines

We provide a detailed description of the guidelines shared with annotators for various tasks below:

## F.1 Annotation Guidelines for Visual Object Identification (Object-Centric QA Generation)

This task involves generating question-answer (QA) pairs based on egocentric video event data by leveraging object interactions in different scenes.

### F.1.1 Data

- **Event List**: Chronologically ordered events describing human actions and the objects involved.
- **Object List**: A global list of unique objects present in the events.

Each event consists of:

- **Action Caption**: Describes the action performed.
- **Local Object List**: Objects involved in the action.

### F.1.2 Annotation Steps

1. **Identify the Visual Scene**
   - Infer the most likely environment based on the object list.
   - Ensure coherence with the given objects.
2. **Select and Replace Objects**
   - Choose at least two objects from the global list.
   - Replace them with **logically relevant new objects** not present in the list.
3. **Generate QA Pairs**
   - Identify events where the selected objects appear.
   - Create a **"Yes" answer** question using the original object.
   - Replace the object and create a **"No" answer** question while keeping the action unchanged.

These guidelines ensure high-quality annotations for object-centric visual understanding.

### F.2 Annotation Guidelines for Episodic Information Reasoning

This task involves generating question-answer (QA) pairs based on egocentric video event data by leveraging object interactions and reasoning about the actions performed.

#### F.2.1 Input Data

- **Event List**: Chronologically ordered events describing human actions and the objects involved.
- **Object List**: A global list of unique objects present in the events.

Each event consists of:

- **Action Caption**: Describes the action performed.
- **Local Object List**: Objects involved in the action.

#### F.2.2 Annotation Steps

1. **Identify the Visual Scene**
   - Infer the most likely environment based on the object list.
   - Ensure coherence with the given objects.
2. **Select and Replace Objects**
   - Choose at least two objects from the global list.
   - Replace them with **logically relevant new objects** not present in the list.
3. **Generate How, Why, or Where Questions**
   - Identify an event containing the selected objects.
   - Select a question type (**How, Why, or Where**) based on the event's nature:
     - If the event describes a **process**, choose a "How" question.
     - If the event describes **reasoning**, choose a "Why" question.
     - If the event describes a **location**, choose a "Where" question.
   - Generate a corresponding question-answer pair.
   - If an event with the new object does not exist, state that the action was not performed.

These guidelines ensure high-quality annotations for episodic information reasoning in egocentric videos.

### F.3 Annotation Guidelines for Temporal Reasoning

This task involves generating question-answer (QA) pairs that require reasoning about the temporal sequence of events in an egocentric video.

#### F.3.1 Input Data

- **Event List**: A chronologically ordered sequence of unique events describing human actions.

Each event consists of:

- **Action Caption**: A description of the action performed.

#### F.3.2 Annotation Steps

1. **Selecting Events from the Event List**
   - Randomly select two events from the chronological list.
   - Ensure that there is a sufficient gap (ideally 4 to 5 events apart).
   - The order should not be directly inferable without examining the full sequence.
2. **Creating Question-Answer Pairs**
   - Formulate questions using the selected events that require reasoning about temporal order.
   - Use words like **"before" and "after"** to indicate event sequencing.
   - Ensure the questions are concise and clear.
   - Generate a corresponding answer based on the event list.

These guidelines ensure high-quality annotations for temporal reasoning in egocentric videos

### F.4 Annotation Guidelines for Object State Change Detection

This task involves identifying and categorizing event sequences from egocentric video data into **complete** and **incomplete** actions, followed by generating corresponding question-answer pairs.

#### F.4.1 Input Data

- **Event List**: A chronologically ordered sequence of unique events describing human actions.

Each event consists of:

- **Action Caption**: A description of the action performed.

#### F.4.2 Annotation Steps

1. **Identifying Complete and Incomplete Actions**
    - **Complete Actions:** A sequence of actions where the object's final state matches its initial state.
    - **Incomplete Actions:** A sequence of actions where the object's final state differs from its initial state.
    - Identify and pair events that meet the above criteria.
2. **Generating Question-Answer Pairs**
    - Formulate questions that require identifying whether an action was completed or left incomplete.
    - Ensure questions are clearly structured and answerable based on the event list.
    - Provide a reasoning statement for each answer.

These guidelines ensure accurate extraction and classification of episodic actions for effective information retrieval.

### F.5 Annotation Guidelines for Visual Object Identification (Action-Centric)

This task involves generating complex question-answer (QA) pairs based on egocentric video event data. The questions should focus on the presence of objects in the activity the person is performing.

#### F.5.1 Input Data

- **Event List**: A chronologically ordered sequence of unique events describing human actions and interactions with objects.
- **Object List**: A global list of unique objects present in the events.
- **Visual Caption**: A description of the most likely activity the person is performing.

Each event consists of:

- **Action Caption**: Describes the action performed.
- **Local Object List**: Objects involved in the action.

#### F.5.2 Annotation Steps

1. **Identify the Activity**
    - Use the visual caption to infer the most likely activity the person is performing.
2. **Select and Replace Objects**
    - Choose at least two objects from the global list.
    - Replace them with **logically relevant new objects** not present in the list.
3. **Generate Question-Answer Pairs**
    - Use the previously identified activity and selected objects to generate questions about whether the person used the object while performing the activity.
    - Ensure that questions align with the event details.
    - Provide a reasoning statement for each answer.

These guidelines ensure accurate question generation for action-centric object identification in egocentric videos.

### F.6 Annotation Guidelines for Hand-Object Interaction

This task involves generating question-answer (QA) pairs to assess fine-grained understanding of human actions by distinguishing between actual and contrastive actions in an egocentric video.

#### F.6.1 Input Data

- **Event List**: A chronologically ordered sequence of unique events describing human actions.

Each event consists of:

- **Action Caption**: A description of the action performed.

#### F.6.2 Annotation Steps

1. **Identify Action Pairs**
    - Randomly select two distinct actions from the event list that describe either:
        - **Physical interaction**: e.g., *"C picks up an object"*, *"C places an object on the table"*.
        - **Movement-based action**: e.g., *"C walks towards the fridge"*.
    - Ensure a gap of **3-5 events** between selected actions to prevent trivial answers.
    - Create contrastive action pairs that invert or contradict the original actions:
        - **Physical interaction contrast**: If the original action is *"C picks up an object"*, the contrast could be *"C throws the object"*.
        - **Movement contrast**: If the original action is *"C walks towards the fridge"*, the contrast could be *"C walks away from the fridge"*.
2. **Generate Question-Answer Pairs**
    - Formulate four QA pairs:
        - Two questions for the original actions (**answer: Yes**).
        - Two questions for the contrastive actions (**answer: No**).

These guidelines ensure accurate annotation of hand-object interactions for assessing action recognition in egocentric videos.

### F.7 Annotation Guidelines for Audio Event Generation

This task involves identifying events in egocentric video sequences where a synthetic background sound can be added. The goal is to introduce plausible ambient sounds that were not originally present but fit within the visual scene.

#### F.7.1 Input Data

- **Event List**: A chronologically ordered sequence of unique events describing human actions.
- **Visual Caption**: A description of the overall activity and environment where the events take place.

Each event consists of:

- **Action Caption**: A description of the action performed.

#### F.7.2 Annotation Steps

1. **Identify Suitable Events**
    - **Filter out events with strong inherent sounds** – If an event naturally produces a dominant sound (e.g., *"C is frying something"* $\rightarrow$ sizzling), avoid adding another cooking-related sound.
    - **Select events where plausible background sounds could occur** – Ensure the sound aligns with the environment and does not contradict the event.
2. **Assign Synthetic Sounds**
    - Choose a background sound that fits the scene but is not naturally produced by the selected event.
    - Ensure the sound is plausible given the visual environment.
    - Avoid contradictions, such as adding an indoor noise in an outdoor setting.

These guidelines ensure high-quality annotations for introducing synthetic background sounds in egocentric videos.

## G Data Source and Filtering

Our dataset was curated primarily from two sources: the VideoRecap (Islam et al., 2024) and Ego4D (Grauman et al., 2022) datasets. Due to inherent challenges within these datasets, specific filtering strategies were employed:

- **Noise Reduction:** Original datasets contain numerous irrelevant or passive scenes. Thus, scenes depicting active interactions with objects were explicitly identified and selected.
- **Static Object Annotation:** To improve model interpretability and rigorously assess recognition capability, all static (non-moving) objects within scenes were carefully annotated using VLLMs.
- **Partial Visibility:** Scenes were specifically chosen where objects were partially obscured or occluded. This intentional selection increases the task complexity and potential for model hallucination.
- **Diverse Task Sampling:** The final dataset includes a wide range of tasks to ensure robustness and generalization in model evaluations.

Given the open-ended response below, determine if the response implies the presence of a visual entity (e.g., character, object, or feature from a digital/virtual world) in an image. The response may include a location or context related to the visual entity. If the response suggests the presence of a visual entity, return "yes". If the response does not imply such a presence, return "no".

Response: <Response>
Virtual Entity: <object>

Return “Yes” or “No”

Figure 7: Details on LLM-as-Judge Prompt

## H LLM-as-Judge Prompt

We have provided details on the LLM-as-Judge prompt used for evaluating open-ended questions in Fig. 7.

## I Additional Details: Auxiliary

**Compute Infrastructure:** All our experiments are conducted on ten NVIDIA A6000 GPUs. No training is required, and depending on the downstream task, a single inference run on a benchmark requires anywhere between 1 to 2 hours.

**Potential Risks:** We manually create all the prompts used in our benchmark to avoid any potential harm or biases.

## J Error Analysis

Apart from hallucinations, we observe the following types of errors in the incorrect responses from MLLMs

**Spatial errors** occur when MLLMs misinterpret the spatial relationships of or among objects, or confuse spatial attributes (where) with temporal or other attributes. These typically occur in "where" questions, with models often providing when, how, or contextual information instead of location.

**Factual errors** occur when MLLMs make mistakes about objective information or details presented in the content. These can appear across various question types but often involve misrepresenting what is actually shown or presented.

**Procedural errors** occur when MLLMs make mistakes in describing the sequence of actions or steps taken. These typically occur in "how" questions, with models often missing important actions or using similar actions but performed in a different context or with a different object.

**Logical errors** occur when answers lack proper reasoning or make invalid inferences that extend beyond what can be supported by the input content. These errors often occur due to the model getting confused in either understanding the difference between after/before or in "why" questions, where models may attribute motives, intentions, or broader contexts that aren’t actually demonstrated in the content.

**No Answer** occurs when the model refuses to give an answer to a question.

| Task | Hallucination Type | Modified Question | Correct Answer | Response |
|---|---|---|---|---|
| Object State Change Detection | Adversarial Sampling | Did the person climb on the bicycle and then recycle it? | Yes | No |
| | | Does the person return the dried grasses to the pile after holding them? | Yes | No |
| Episodic Information Reasoning | Prompt Injection | Where did the person drop the water hose? | The person dropped the water hose on the staircase. | The person is walking back into the garage, past some equipment, and the water hose drops at [03:15]. |
| | | How did the person collect the nail? | The person did not perform this action. | The person collects a small nail that was in the fuel injector of the machine using a magnet. |
| Hand-Object Interaction | Adversarial Sampling | Did the person put the mixture in a frying pan? | Yes | No. The person scoops a mixture with a spoon and places it onto a plate. There is no frying pan in the video. |
| | | Did the person throw the pair of scissors? | No | Yes, it appears the person threw the scissors aside on the floor. |
| Visual Object Identification | Adversarial Sampling | Did the man take a box on the table? | Yes | No, he didn't. |
| | | Did the person wipe the oil tank with the piece of sponge? | No | Yes, the person picked up a piece of sponge and wiped the oil tank with it. |

Table 5: Few examples of failure cases for Gemini.