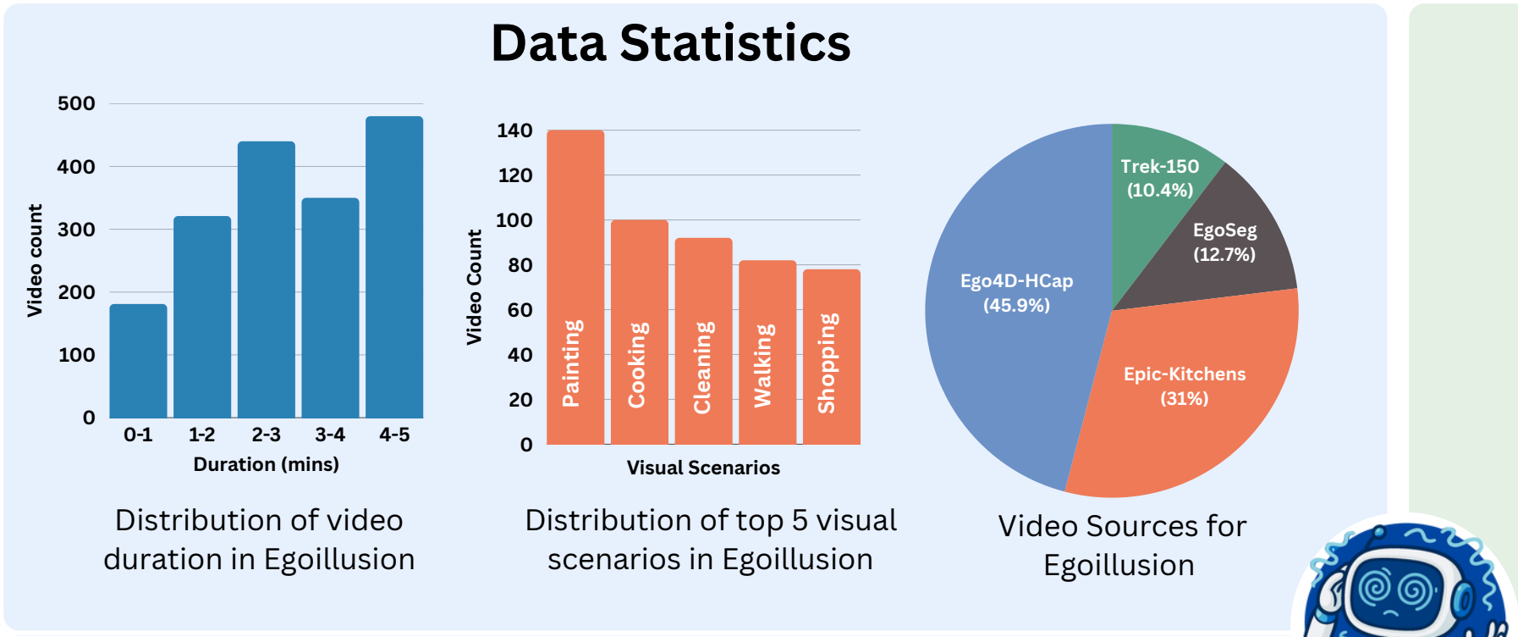
EgoIllusion: Benchmarking Hallucinations in Egocentric Video Understanding

Ashish Seth[1]*, Utkarsh Tyagi[1]*, Ramaneswaran Selvakumar[1], Nishit Anand[1], Sonal Kumar[1], Sreyan Ghosh[1], Ramani Duraiswami[1], Chirag Agarwal[2], Dinesh Manocha[1]

[1]University of Maryland, College Park [2]University of Virginia

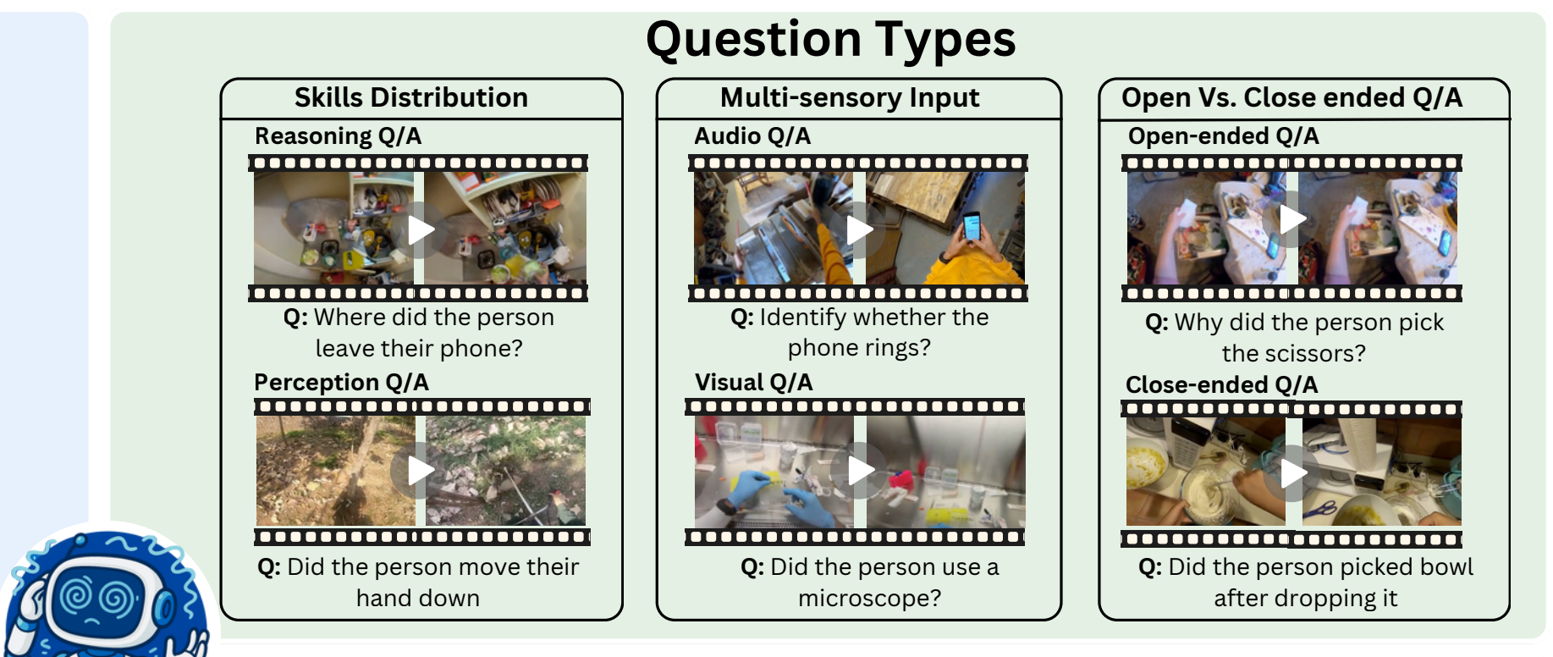
Introduction


Data Statistics

Distribution of video duration in Egoillusion

Distribution of top 5 visual scenarios in Egoillusion

Video Sources for Egoillusion


Question Types

Q: Where did the person leave their phone?

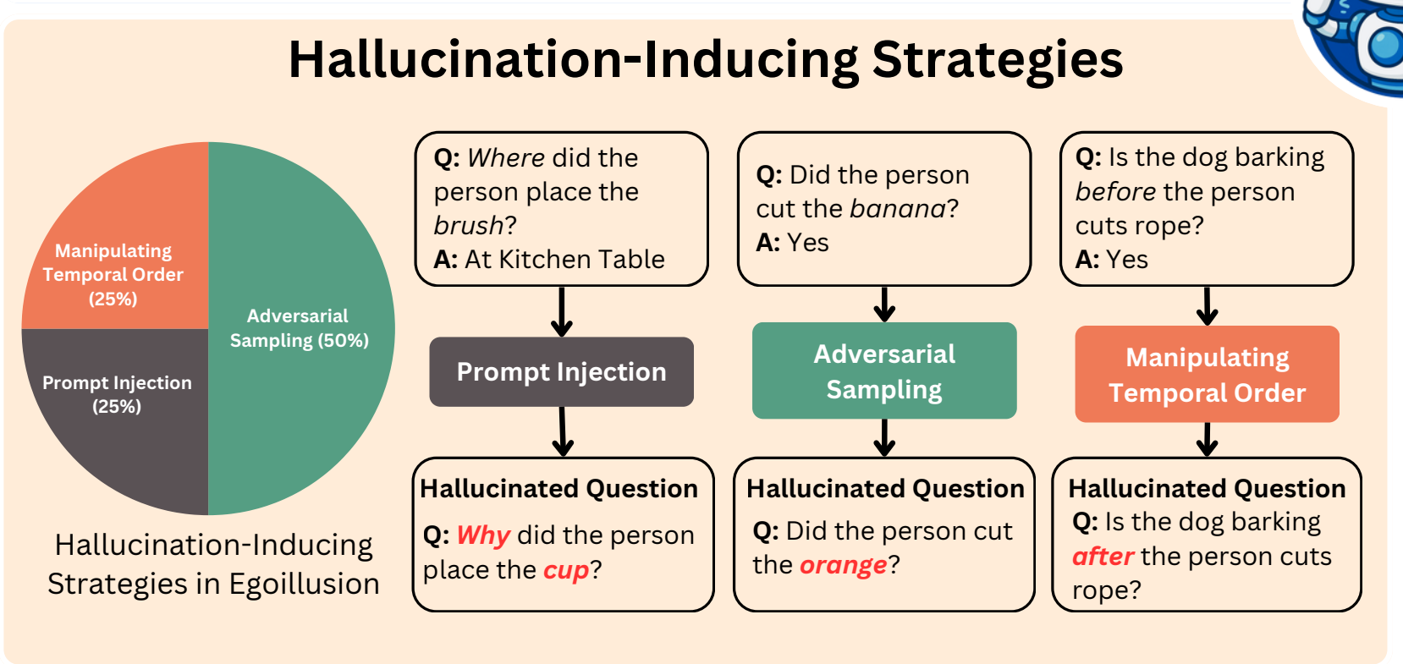
Q: Did the person move their hand down

Q: Identify whether the phone rings?

Q: Did the person use a microscope?

Q: Why did the person pick the scissors?

Q: Did the person picked bowl after dropping it


Hallucination-Inducing Strategies
Manipulating Temporal Order (25%)
Adversarial Sampling (50%)
Prompt Injection (25%)
Prompt Injection
Adversarial Sampling
Manipulating Temporal Order
Hallucinated Question

Q: *Where* did the person place the *brush*?
A: At Kitchen Table

Q: Did the person cut the *banana*?
A: Yes

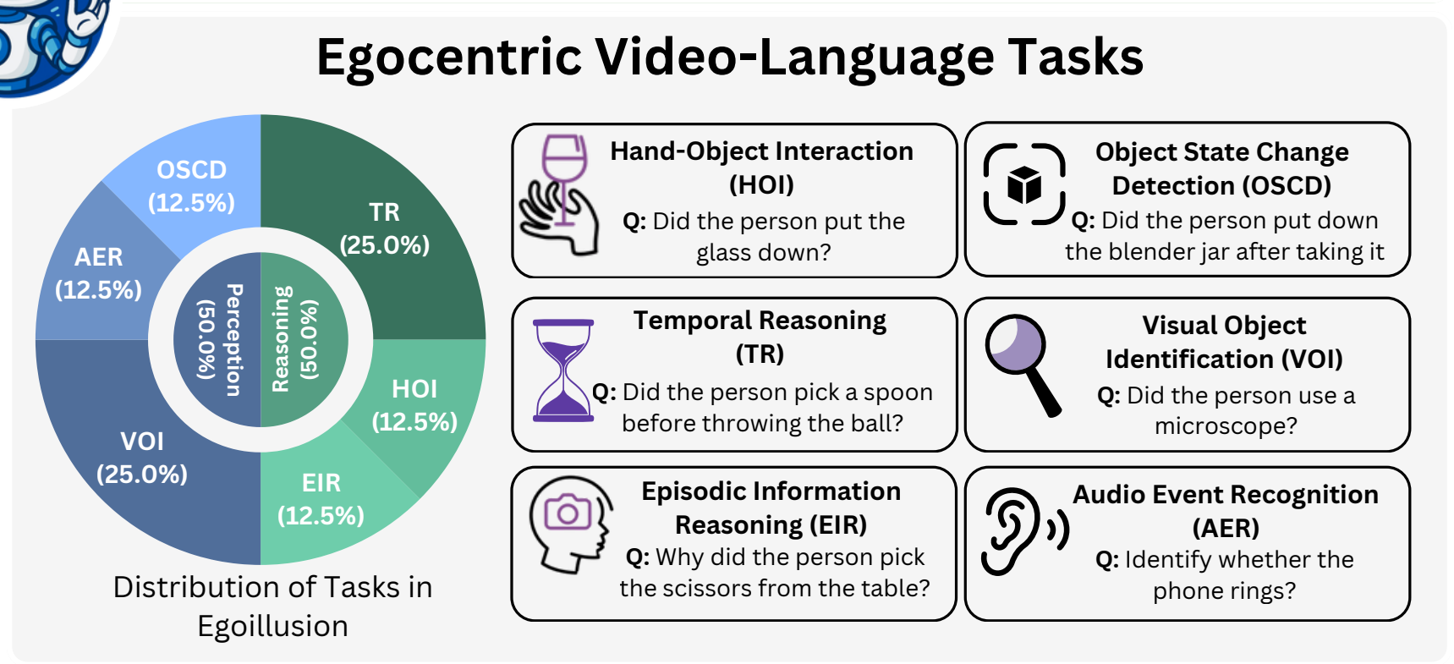
Q: Is the dog barking *before* the person cuts rope?
A: Yes

Hallucinated Question
Q: ***Why*** did the person place the ***cup***?

Hallucinated Question
Q: Did the person cut the ***orange***?

Hallucinated Question
Q: Is the dog barking ***after*** the person cuts rope?

Hallucination-Inducing Strategies in Egoillusion


Egocentric Video-Language Tasks

Hand-Object Interaction (HOI)
Q: Did the person put the glass down?

Object State Change Detection (OSCD)
Q: Did the person put down the blender jar after taking it

Temporal Reasoning (TR)
Q: Did the person pick a spoon before throwing the ball?

Visual Object Identification (VOI)
Q: Did the person use a microscope?

Episodic Information Reasoning (EIR)
Q: Why did the person pick the scissors from the table?

Audio Event Recognition (AER)
Q: Identify whether the phone rings?

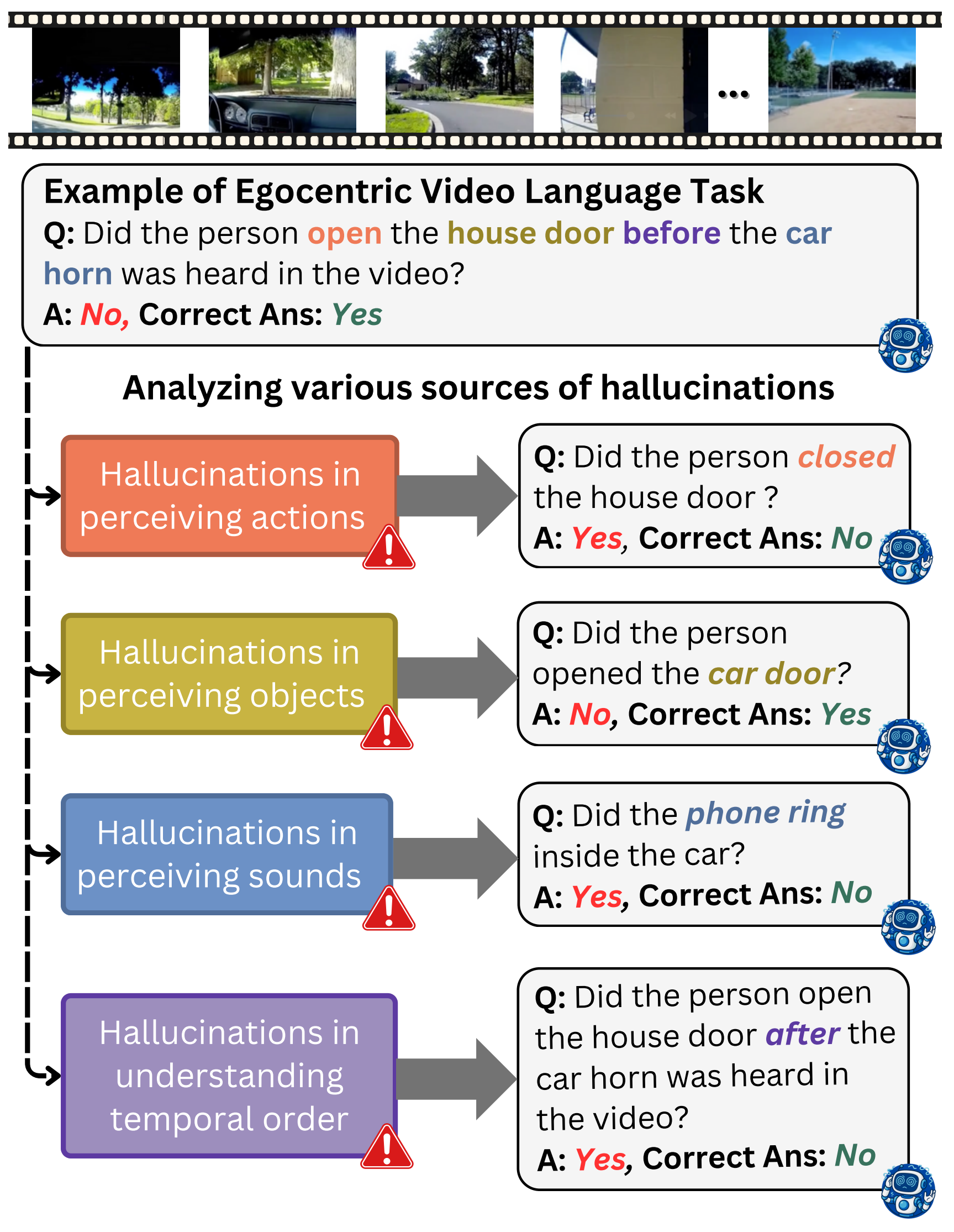
Distribution of Tasks in Egoillusion

Our 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:

- **Perception vs Reasoning:** distinguishing between perceptual and reasoning skills by evaluating object recognition, action understanding, and scene inference
- **Multisensory Inputs:** integrating visual and auditory cues, such as object appearance, human actions, and environmental sounds, to assess multimodal alignment
- **Question Types:** supporting both closed-ended and open-ended questions, requiring models to answer factually grounded queries while reasoning about events and interactions.

Motivation

Egocentric videos captured from wearable devices primarily capture human-object interactions, providing rich multisensory information. Although MLLMs demonstrate strong performance on standard image and video benchmarks, they remain susceptible to hallucinations, producing coherent but incorrect interpretations of sensory input that diverge from reality. As illustrated in the figure, state-of-the-art MLLMs such as Gemini 1.5 Pro exhibit a high 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.*


Example of Egocentric Video Language Task
Q: Did the person open the house door before the car horn was heard in the video?
A: *No,* **Correct Ans:** *Yes*

Analyzing various sources of hallucinations

Hallucinations in perceiving actions
Hallucinations in perceiving objects
Hallucinations in perceiving sounds
Hallucinations in understanding temporal order

Q: Did the person *closed* the house door ?
A: *Yes*, **Correct Ans:** *No*

Q: Did the person opened the *car door*?
A: *No*, **Correct Ans:** *Yes*

Q: Did the *phone ring* inside the car?
A: *Yes*, **Correct Ans:** *No*

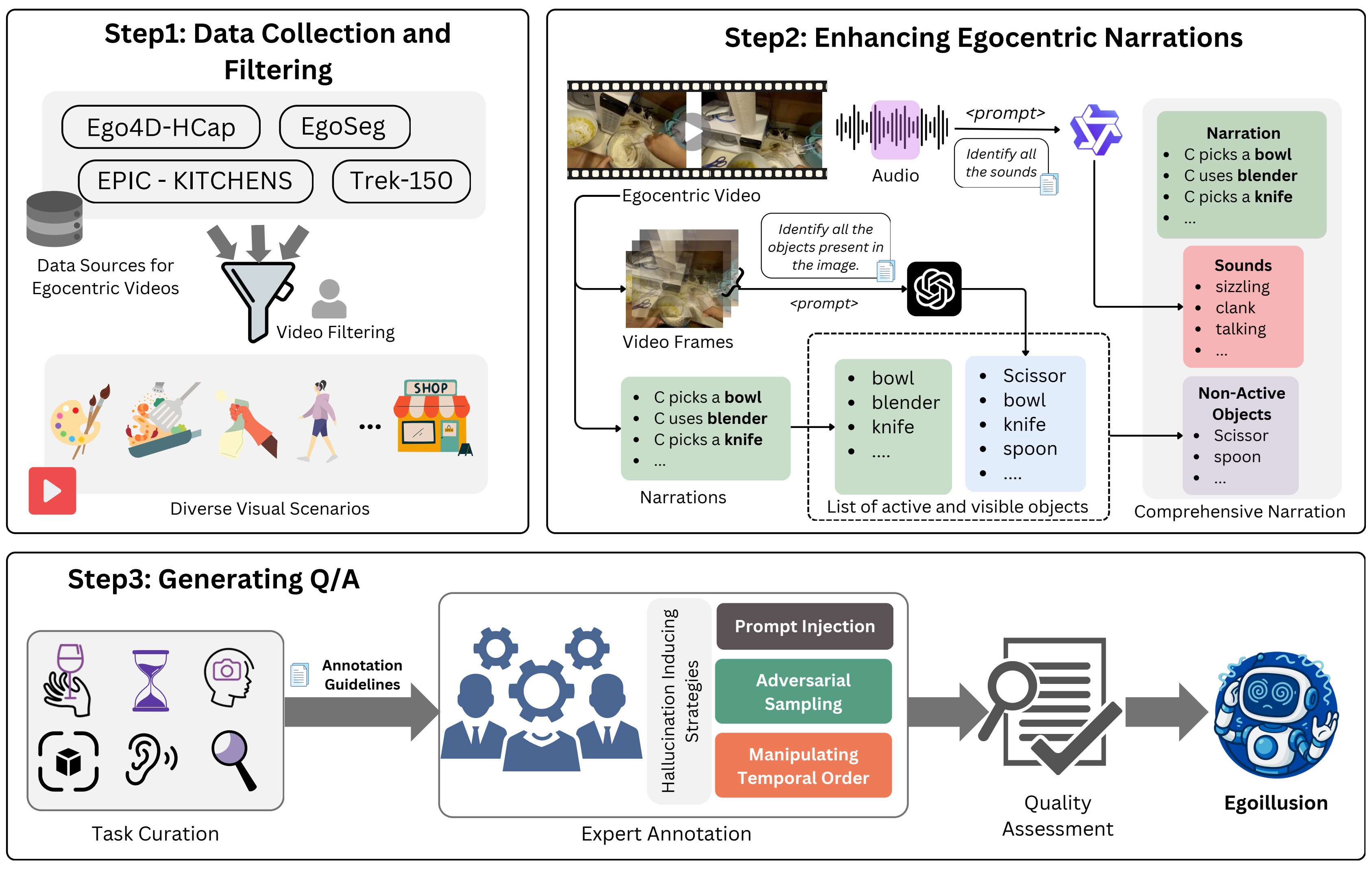
Q: Did the person open the house door *after* the car horn was heard in the video?
A: *Yes*, **Correct Ans:** *No*

EgoIllusion Vs. Prior Benchmarks

Benchmark	Size	Modality		Skills	
		Vision	Audio	Perception	Reasoning
POPE	3k	✓	×	3k ✓	0 ×
HallusionBench	1.1k	✓	×	0 ×	1.1k ✓
MMHal-Bench	0.1k	✓	×	0.05k ✓	0.05k ✓
Bingo	0.4k	✓	×	0 ×	0.4k ✓
EasyDetect	0.4k	✓	×	0.4k ✓	0 ×
VHTest	1.2k	✓	×	0.6K ✓	0.6K ✓
VALOR	0.2k	✓	×	0.2k ✓	0 ×
VideoHallucer	1.8k	✓	×	0.9k ✓	0.9k ✓
EgoIllusion *(ours)*	8k	✓	✓	4.0k ✓	4.0k ✓

We compare our benchmark EgoIllusion with existing multimodal hallucination benchmarks. In contrast to other benchmarks, EgoIllusion covers both vision and audio modality, while having the highest number of perception and reasoning-based questions.

Data Construction Pipeline



We first collect egocentric videos with detailed narrations from open-source datasets like Ego4D-HCap and EPIC-KITCHENS, 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 and incorporating environmental sounds via Qwen2-Audio. 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.

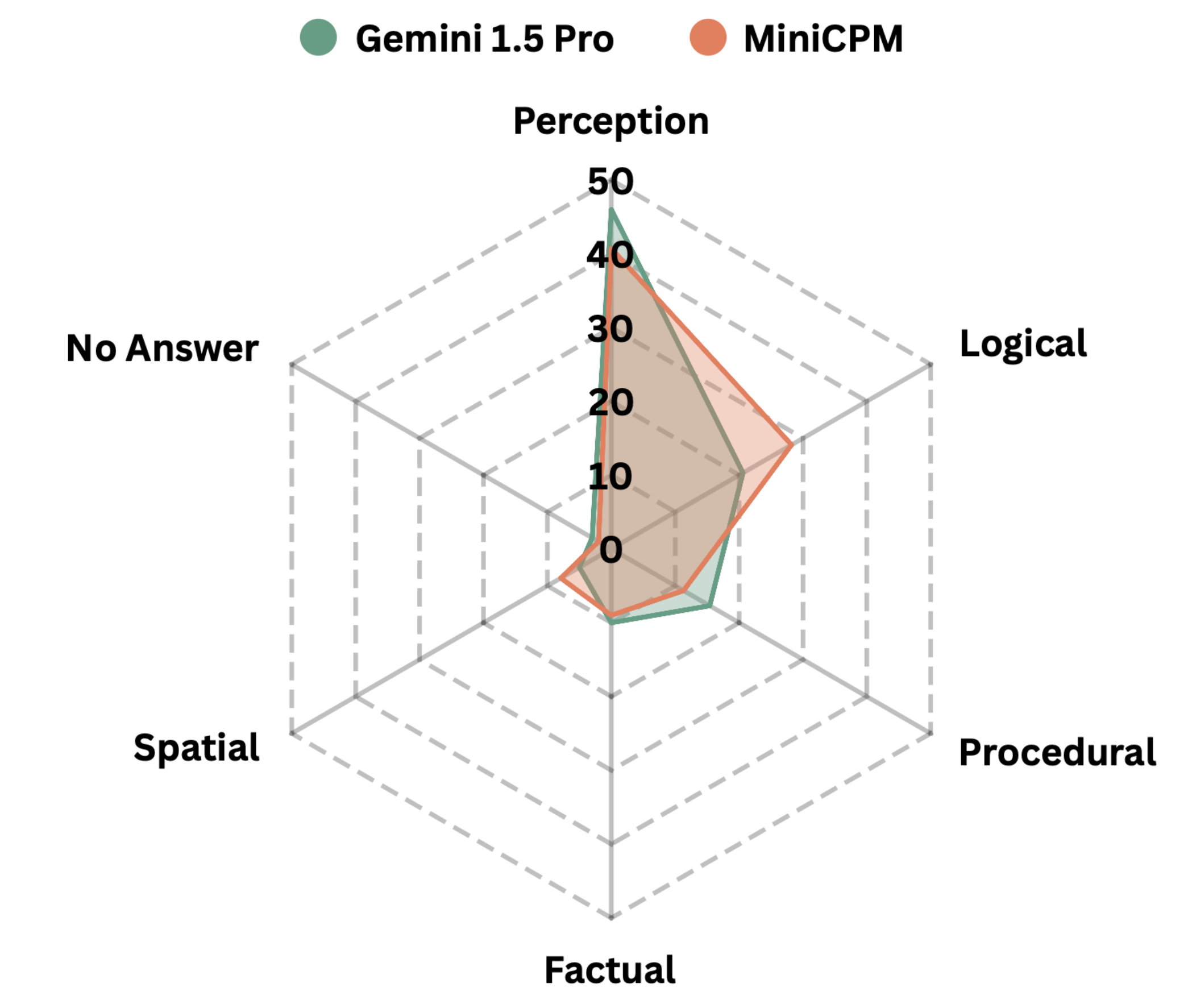
Results

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	3B	×	✓	×	50.1$_{\pm0.3}$	67.3$_{\pm0.2}$	54.6$_{\pm0.4}$	56.3$_{\pm0.1}$	51.1$_{\pm0.3}$	-	55.8$_{\pm0.2}$
VideoLlama3	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	8B	✓	✓	×	51.4$_{\pm0.4}$	64.3$_{\pm0.1}$	65.5$_{\pm0.2}$	60.8$_{\pm0.3}$	51.7$_{\pm0.2}$	-	58.7$_{\pm0.3}$
LLaVa-NEXT	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	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	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	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	8B	×	✓	✓	**57.3$_{\pm0.4}$**	47.3$_{\pm0.1}$	**66.9$_{\pm0.5}$**	69.5$_{\pm0.3}$	58.4$_{\pm0.2}$	50.1$_{\pm0.4}$	58.9$_{\pm0.3}$
VideoLlama2	7B	✓	✓	✓	56.1$_{\pm0.3}$	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	-	-	✓	✓	51.4$_{\pm0.2}$	60.8$_{\pm0.3}$	61.8$_{\pm0.5}$	68.1$_{\pm0.4}$	56.5$_{\pm0.1}$	52.5$_{\pm0.3}$	**59.4$_{\pm0.2}$**
GPT-4o	-	-	✓	×	53.2$_{\pm0.3}$	47.5$_{\pm0.2}$	66.7$_{\pm0.4}$	**73.9$_{\pm0.5}$**	58.4$_{\pm0.1}$	-	58.8$_{\pm0.3}$

We compare the performance 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 underlined. *Overall, most of the MLLMs show performance close to the random guess on our benchmark.*

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. 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.*


Gemini 1.5 Pro
MiniCPM
Perception
Logical
Procedural
Factual
Spatial
No Answer