EgoSound: Benchmarking Sound Understanding in Egocentric Videos

Bingwen Zhu[1,2*] Yuqian Fu[1,3,6*†] Qiaole Dong[1] Guolei Sun[4] Tianwen Qian[5]
Yuzheng Wu[1] Danda Pani Paudel[3] Yanwei Fu[1,2] Xiangyang Xue[1]
[1]Fudan University [2]Shanghai Innovation Institute [3]INSAIT, Sofia University "St. Kliment Ohridski"
[4]Nankai University [5]East China Normal University [6]KAUST

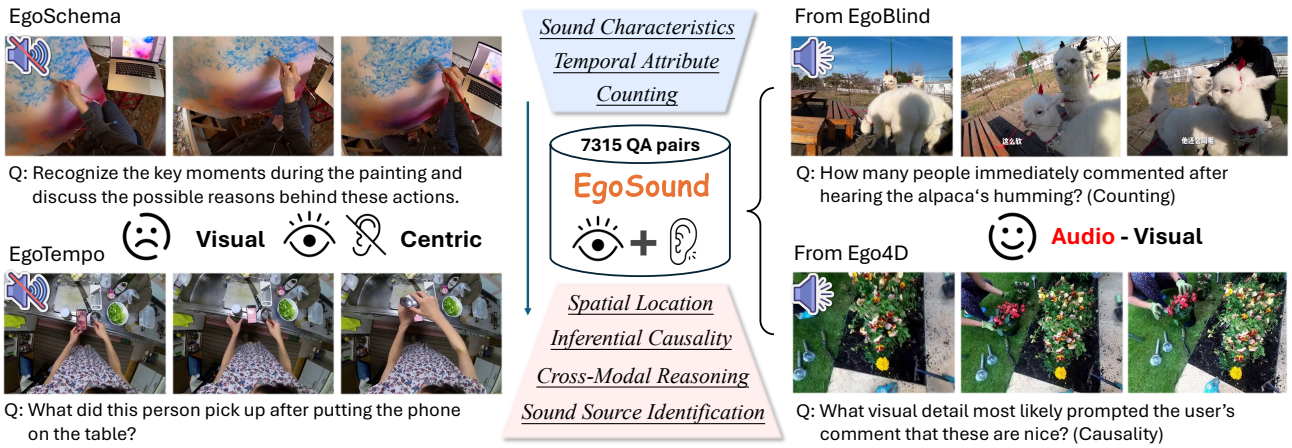

Figure 1. **EgoSound *vs* existing egocentric Video Question Answering (VideoQA).** Prior datasets (left) [23, 27] focus solely on vision-centric question answering with no awareness of audio, whereas EgoSound (right) constructs a more complex and comprehensive audio-visual QA dataset tailored for sound understanding. It is built from two dataset sources [13, 36], includes 900 videos and 7315 high-quality QA pairs, and spans seven task categories—making it a benchmark that can both listen and see.

Abstract

Multimodal Large Language Models (MLLMs) have recently achieved remarkable progress in vision-language understanding. Yet, human perception is inherently multisensory, integrating sight, sound, and motion to reason about the world. Among these modalities, sound provides indispensable cues about spatial layout, off-screen events, and causal interactions, particularly in egocentric settings where auditory and visual signals are tightly coupled. To this end, we introduce EgoSound, the first benchmark designed to systematically evaluate egocentric sound understanding in MLLMs. EgoSound unifies data from Ego4D and EgoBlind, encompassing both sighted and sound-dependent experiences. It defines a seven-task taxonomy spanning intrinsic sound perception, spatial localization, causal inference, and cross-modal reasoning. Constructed through a multi-stage auto-generative pipeline, EgoSound contains 7315 validated QA pairs across 900 videos. Comprehensive experiments on nine state-of-the-art MLLMs reveal that current models exhibit emerging auditory reasoning abilities but remain limited in fine-grained spatial and causal understanding. EgoSound establishes a challenging foundation for advancing multisensory egocentric intelligence, bridging the gap between seeing and truly hearing the world. Project page: https://groolegend.github.io/EgoSound/.

1. Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable progress in integrating vision and language, enabling sophisticated visual understanding and reasoning. Yet, true human-like perception extends far beyond vision—it is inherently multisensory, grounded in the seamless integration of sound, touch, and motion. Among these, sound plays a particularly vital role: it conveys spatial cues, reveals off-screen events, and en-

*Equal Contribution.
†Corresponding Author.

codes the causality and intent behind interactions. This becomes especially critical in egocentric settings, where the auditory and visual streams are deeply intertwined, capturing the world as directly experienced by the wearer.

Despite this importance, research in egocentric perception has remained overwhelmingly visual-centric. Most prior works [4, 9–12, 15, 18–22, 25, 27, 29, 35, 36, 42, 44, 47] focus on recognizing and predicting events visible in the scene, while treating audio as secondary—or ignoring it entirely. Perceiving the world egocentrically without sound is akin to navigating a silent world, fundamentally limiting the depth of understanding. For instance, the sharp hiss of steam or the sudden clatter of metal carries vital information that vision alone may miss. Likewise, for individuals with visual impairments, sound is not auxiliary but essential for navigation and situational awareness.

To fill this gap, we introduce **EgoSound**, a new benchmark that systematically evaluates the egocentric sound understanding capabilities of MLLMs [5, 7, 31, 37, 38, 41]. EgoSound is the first dataset explicitly designed to study nuanced audio-visual reasoning from a first-person perspective, encompassing both environmental sounds from human-object interactions and human dialogues that drive contextual understanding. Its mission is simple yet profound: to enable models that can hear, not just see, from a first-person viewpoint. A distinguishing feature of EgoSound is its multi-source design, which integrates videos from both the large-scale Ego4D dataset [13], capturing a wide range of everyday activities, and the EgoBlind dataset [36], focusing on scenarios where auditory perception is essential for understanding, interaction, and navigation. This combination provides a comprehensive coverage of egocentric experiences, spanning from visually guided to sound-dependent contexts.

We further propose a novel taxonomy of seven egocentric sound tasks that span both unimodal and multimodal reasoning—from intrinsic sound properties (e.g., Sound Characteristics, Counting, Temporal Attribute) to complex audio-visual reasoning (e.g., Spatial Location, Source Identification, Inferential Causality, Cross-Modal Reasoning). To construct EgoSound, we develop a multi-stage data curation pipeline leveraging modern generative models (Qwen2.5-VL [2], Gemini-2.5 [7], GPT-4o [17]). The pipeline first identifies key human-object interactions, then generates rich audio-centric captions, and finally constructs high-quality, open-ended question–answer pairs focused on sound-related reasoning. The final dataset comprises 7315 validated Q&A pairs over 900 rigorously filtered videos, ensuring strong fidelity and task diversity.

We evaluate eight state-of-the-art MLLMs, including models from the Qwen-Omni [37, 38], video-SALMONN 2+ [31], VideoLLaMA2.1 [5], and MiniCPM [41] families, as well as the egocentric-specialized EgoGPT [40]. Our comprehensive experiments reveal that while current MLLMs exhibit emerging auditory reasoning abilities, they still struggle with fine-grained spatial, temporal, and causal inference based on sound. EgoSound thus establishes a challenging, high-quality benchmark for multisensory egocentric intelligence—bridging a critical gap in current MLLMs and paving the way for future research toward models that can listen, understand, and reason about the full multisensory world.

Our main contributions are summarized as follows:

- **EgoSound:** The first large-scale benchmark for egocentric sound understanding in MLLMs, featuring data from both sighted (Ego4D) and blind (EgoBlind) perspectives.
- **A Novel Task Taxonomy:** Seven tasks covering intrinsic sound perception, spatial reasoning, causal inference, and cross-modal understanding.
- **High-Quality Dataset:** 7315 validated open-ended Q&A pairs created via a rigorous, multi-stage curation pipeline centered on sound events.
- **Comprehensive Benchmarking:** Evaluation of nine cutting-edge MLLMs, uncovering key challenges in egocentric audio-visual reasoning and establishing strong baselines for future research.

2. Related Work

Egocentric Video Question Answering. To advance video question answering (VideoQA) on egocentric videos, numerous datasets [8, 14, 16, 26, 34, 36] and benchmarks [4, 9, 12, 18, 20, 23, 27, 29, 39, 40, 44] have been proposed. Early and representative ones include EgoVQA [9], EgoTaskQA [18], and EgoSchema [23], each focusing on different aspects and characteristics of first-person understanding, covering a range of tasks such as descriptive, predictive, explanatory, and counterfactual reasoning. More recently, several complementary benchmarks have also been introduced, broadening the evaluation scope. For example, EgoThink [4] assesses models' ability to "think" from a first-person perspective across multiple reasoning dimensions; AMEGO [12] emphasizes long-term temporal reasoning and memory over extended egocentric videos; EgoTempo [27] focuses on temporal understanding through long-horizon questions; EASG-Bench [29] introduces scene-graph-based QA to capture spatio-temporal relations; EgoCross [20] explores cross-domain generalization across diverse first-person scenarios such as surgery and sports; Despite covering diverse visual and cognitive skills, most existing egocentric QA benchmarks rely solely on visual cues, overlooking the rich auditory context in first-person videos. In contrast, EgoSound focuses on auditory-(visual) cues to promote a more comprehensive understanding of egocentric scenes. Tab. 1 provides a multidimensional comparison between EgoSound and prior egocentric QA benchmarks.

Dataset	Video Length	Clips	QA Pairs	Categories	Sound Questions	Multiple Sources	Open-ended
EgoVQA [9]	(25, 100)s	520	0.6k	5	✗	✗	✓
EgoTaskQA [18]	25s	2336	40k	4	✗	✗	✓
EgoSchema [23]	3min	1981	5k	-	✗	✗	✗
EgoThink [4]	-	-	0.75k	6	✗	✗	✓
AMEGO [12]	14min	100	20.5k	8	✗	✗	✗
EgoTempo [27]	45s	365	0.5k	10	✗	✗	✓
EASG-Bench [29]	3.1min	221	1.8k	5	✗	✗	✓
EgoCross [20]	22.5s	798	0.95k	4	✗	✓	✓
EgoSound (ours)	59s	900	7.3k	7	✓	✓	✓

Table 1. Comparison with existing egocentric video QA benchmarks. EgoSound is distinguished by its focus on sound-centric reasoning.

Audio(-Visual) Question Answering. Audio-based Question Answering (Audio-QA) aims to reason about spatial and semantic information from audio or audio–visual observations to answer natural language queries. Early research on audio reasoning primarily focused on localization and detection rather than QA, such as STARSSS23 [30]. Subsequent work shifted toward question answering with explicit spatial reasoning benchmarks. SpatialSoundQA [45] and AQAPHY [33] introduced different benchmark for Audio-QA, with distinct focuses of reasoning tasks. SAVVY [3] and Magnet [6] further extend to audio–visual QA, emphasizing the joint understanding of auditory and visual modalities. Despite these advances, Audio(-Visual) QA in the *egocentric* domain remains largely unexplored, overlooking the fact that audio plays a crucial role in understanding and grounding the world from a first-person perspective.

Multimodal Large Language Models. Multimodal large language models (MLLMs) have advanced rapidly, with a surge of commercial systems (e.g., GPT-4 [1], Gemini 2.5 Pro [7]) and open-source counterparts (e.g., Qwen2.5-VL [2], InternVL [48], Video-LLaMA [5, 43], MiniCPM [41], video-SALMONN [31], and Qwen-Omni [37, 38]), which demonstrate strong reasoning ability across benchmarks including MLVU [46], LVBench [32], and EgoSchema [23]. In the egocentric domain, concurrent progress in both models and datasets has led to several pretrained egocentric MLLMs, such as EgoVLPv2 [28] and xEgoGPT [40], However, *audio*, a key cue for human perception and situational awareness, remains underexplored in both model design and evaluation. Therefore, in this work, we focus on evaluating how MLLMs perceive and reason over sound cues and joint audio–visual cues under a first-person perspective.

3. EgoSound Dataset

We introduce EgoSound, a novel benchmark to systematically evaluating performance of MLLMs across a variety of egocentric sound understanding tasks. Egosound covers both sighted and blind perspectives, and consists of egocentric videos spanning various visual scenarios. In this section, we provide a comprehensive introduction to the EgoSound benchmark. We begin by discussing the video data selection, and the taxonomy of question-answering tasks, followed by an explanation of the data curation pipeline, and conclude with dataset statistics.

3.1. Data Collection and Filtering

The videos included in EgoSound are carefully sourced from egocentric datasets with rich audio, including Ego4D [13], and EgoBlind [36]. This selection ensures a broad spectrum of acoustic environments and contexts. Ego4D [13], as the most extensive egocentric video dataset, contributes a wide variety of scenarios from daily life and work. The inclusion of EgoBlind [36] is particularly unique, offering data recorded by blind individuals where auditory cues are often central to daily navigation and interaction. Collectively, these sources ensure that our dataset covers a wide range of human activities such as sports, learning, doing chores, cooking, commuting, driving, shopping, playing instruments, and filming, across numerous indoor and outdoor environments. The videos in EgoSound span a broad range of durations, from short clips of 5 seconds to extended recordings of up to 5 minutes, capturing both brief atomic sounds and more complex, temporally evolving acoustic events.

To ensure data quality, the construction of EgoSound involved a rigorous filtering process for both audio and visual streams. For the audio, we first discarded videos with extended silence, excessive background noise, or unintelligible speech. The remaining clips were then carefully trimmed to retain segments rich in meaningful sound events. This process concentrates the data on high-quality sound suitable for generating audio-centric question-answer pairs. Visually, we removed clips that were static or monotonous. We specifically retained segments that feature dynamic human activities and rich object interactions. Through this dual filtering strategy, we curated 900 egocentric videos

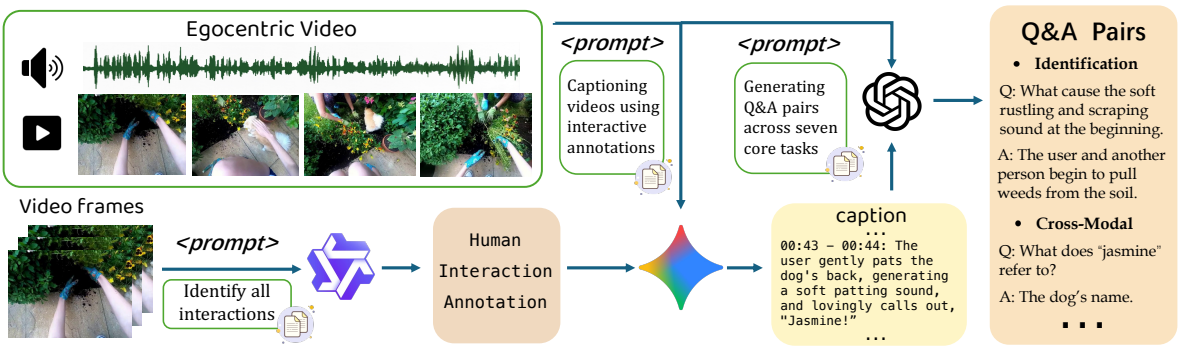

Figure 2. **Overview of the EgoSound data curation pipeline.** We first identifies human interaction events, then generates interaction-grounded and sound-centric audio-visual captions, and finally build visually-verified OpenQA pairs corresponding to the seven core tasks.

featuring rich and complex audio-visual scenarios. The filtering process is crucial for creating challenging and high-quality queries to effectively evaluate the audio-visual understanding capabilities of MLLMs.

3.2. Task Taxonomy

Design Principles. Our taxonomy follows three principles: a) Literature Grounding: We adapt established AudioQA and egocentric VideoQA tasks in Sec. 2, structuring them into sound-dependent vs. audio-visual categories; b) Comprehensive Assessment: We select diverse, complementary categories to ensure holistic evaluation; and c) Practical Relevance: We value tasks central to real-world scenarios aiming to support downstream applications. As illustrated in Fig. 3 (d), we curate seven egocentric sound tasks that target core capabilities essential for audio-visual understanding. These tasks not only investigates the intrinsic properties of sounds, but also explores multimodal perception and reasoning. The intrinsic sound properties involve *Sound Characteristics*, *Counting*, and *Temporal Attribute*, while the multimodal perception and reasoning aspects cover *Spatial Location*, *Sound Source Identification*, *Inferential Causality*, and *Cross-Modal Reasoning*.

- **Sound Characteristics**. These tasks assess models' ability to describe the intrinsic acoustic properties of a sound, such as its perceived volume, texture, or timbre.
- **Counting**. Counting tasks are designed to evaluate models' ability to track and enumerate distinct instances of auditory events or the repetitions of a specific sound. It also includes the number of times a particular word or phrase is mentioned in speech.
- **Temporal Attribute**. Temporal tasks are proposed to evaluate models' ability to analyze the temporal dynamics of a sound, including its duration, specific timing, and how its acoustic features evolve over time.
- **Spatial Location**. Location tasks test models' ability to localize a sound source in three-dimensional space relative to the egocentric observer, identifying both its direction and approximate distance.
- **Sound Source Identification**. Identification tasks evaluate models' ability to identify the specific object or action that produced a sound, requiring the model to ground auditory signals to their corresponding visual events.
- **Inferential Causality**. Causality tasks test models' higher-level ability to reason about the underlying cause or intent behind an auditory event by synthesizing information from the surrounding audio-visual context.
- **Cross-Modal Reasoning**. Cross-Modal tasks assess a models' ability to integrate information across both modalities for complex inference. This includes using audio to interpret visual events (Audio-Guided Visual Reasoning) and using visual context to explain auditory events (Visual-Guided Audio Reasoning).

While the source egocentric video data provide meta annotation [13] or question-answer (Q&A) pairs [36], these annotations are predominantly focused on visual content. Consequently, they are insufficient for our primary goal of facilitating a deep and nuanced evaluation of sound event understanding. To address this gap, we designed a multi-stage data curation pipeline to generate high-quality, audio-centric Q&A Pairs for EgoSound dataset. We illustrate our data curation pipeline in Fig. 2. Our pipeline consists of three key stages: Human Interaction Annotation, Audio-Visual Caption Generation, and Q&A Pairs Construction.

3.3. Data Curation Pipeline

Human Interaction Annotation. Directly prompting an omni model for a general description of an egocentric video often results in captions that overlook significant scene details. Recognizing that physical interactions are the primary source for meaningful sound events, the first stage of our pipeline is designed to systematically annotate these key moments. To achieve this, we leverage the Qwen2.5-VL [2] to perform automated annotation of human-object and human-human interactions. This process generates temporally-grounded labels that capture specific actions

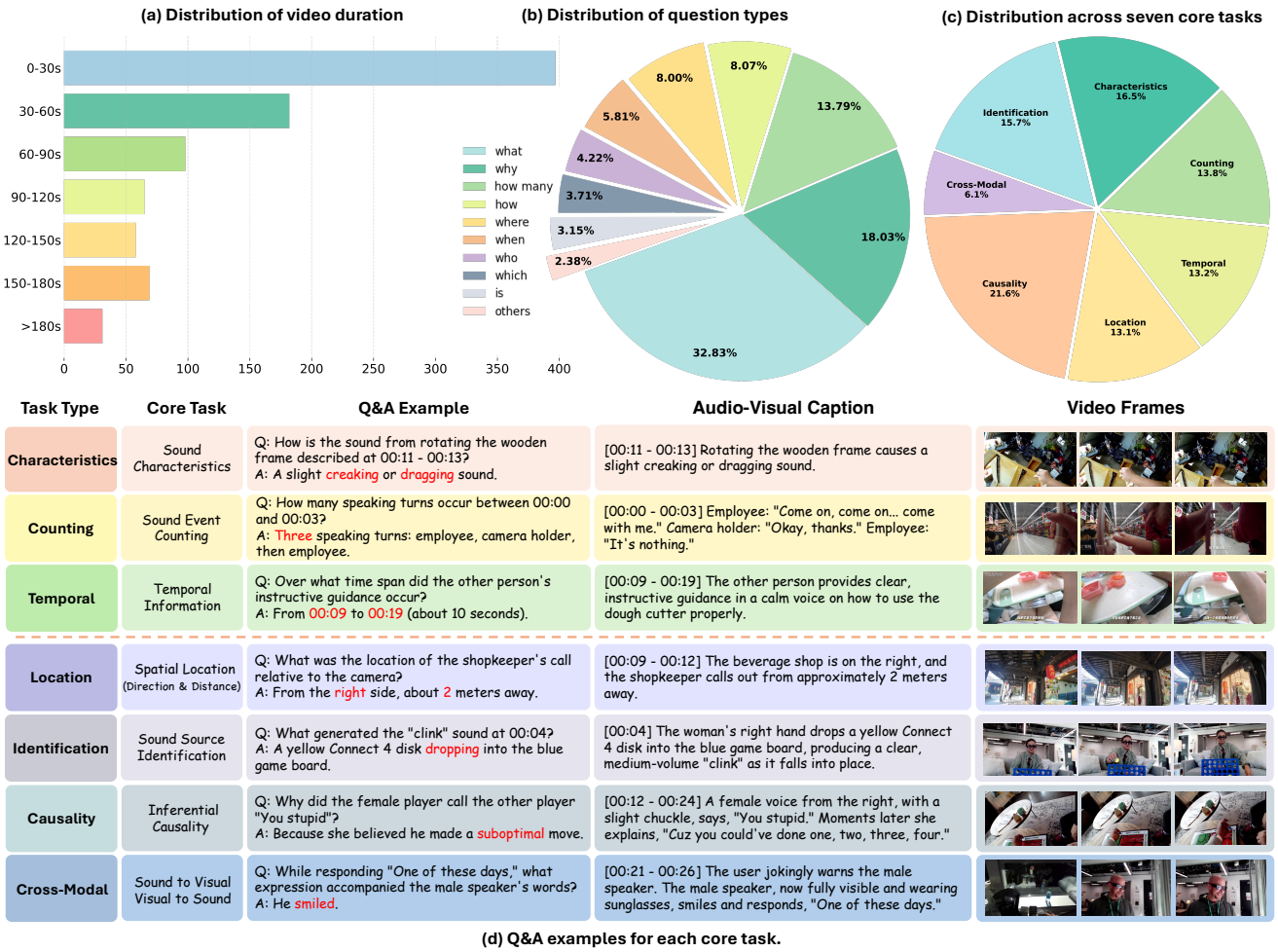

Figure 3. **Overview of the EgoSound task taxonomy and statistics.** (Top) Statistics on video length, question type, and the number of questions for each task category. (Bottom) A selection of representative examples for each core task of EgoSound.

(e.g. "At 3s, a girl in a red dress picks up the camera.", "From 45s to 48s, a man in a white shirt drives a black car past the camera wearer."). These structured, interaction-focused annotations serve as a rich contextual foundation. They are used as conditioning prompts to guide the model in generating comprehensive audio-visual captions in the next stage, ensuring that the descriptions are anchored to the specific events that produce sound.

Audio-Visual Caption Generation. With egocentric sound understanding as the primary goal, the second stage of our pipeline generates audio–visual captions that center on sound while using visual human interaction labels as auxiliary context. Rather than describing the video broadly, we leverage the interaction annotations to disambiguate and refine the interpretation of audio, steering the model across seven sound-centric tasks (Sec. 3.2). Concretely, for each annotated interaction, we prompt Gemini-2.5 [7] to describe the corresponding audio by linking each sound to its source, its acoustic traits, how many sources are active, when it occurs, how long it lasts, where it is in space, why it occurs, and how visual context helps explain audio. Additionally, all spoken words are also transcribed into the caption. Specific prompts are provided in the Supplementary. This yields captions grounded in first-person scenes but optimized for fine-grained, source-aware audio annotation, supporting high-quality audio–visual Q&A in the final stage.

Q&A Pairs Construction. To construct high-quality question–answer pairs while reducing potential hallucination effects in Gemini [7], we instruct GPT-4o [17] to generate meaningful Q&A samples based on detailed audio–visual captions and their corresponding video clip frames. We prompt GPT-4o [17] to ask questions across seven core sound-centric tasks (Sec. 3.2), with answers derived directly from the captions or inferred within their con-

textual bounds. As a double validation, every Q&A pairs should be supported by visual evidence found in the video frames to ensure factual consistency.

To ensure the rigor of the evaluation and prevent guesswork, we adopted an open-ended question-answer (OpenQA) format, requiring the tested model to give descriptive answers rather than selecting from a list of options. Following this pipeline, we constructed a total of 7315 validated QA pairs to comprehensively evaluate the audio-visual comprehension capabilities of MLLMs.

3.4. Dataset Statistics

Tab. 2 summarizes the data sources of EgoSound. The dataset comprises 900 egocentric videos, including 640 clips from EgoBlind and 260 from Ego4D. These videos vary in duration, ranging from 5 seconds to 5 minutes, with an average length of 59 seconds. The overall distribution of video durations is shown in Fig. 3 (a). In total, 7315 QA pairs are included, of which 4,969 are derived from EgoBlind, and the remaining ones from Ego4D. The proportions of QA pairs corresponding to the seven task categories are illustrated in Fig. 3 (c). We further analyze the distribution of question types, as shown in Fig. 3 (b).

Dataset	Clips	QA Pairs	Dur.(s)
EgoBlind [36]	640	4969	40.5
Ego4D [13]	260	2346	105.6
EgoSound	900	7315	59.3

Table 2. Data source statistics for EgoSound.

3.5. Human Verification

From over 7000 QA pairs, we balancedly sampled 350 QA pairs to conducted human verification study, with 50 pairs for each of the seven tasks. To ensure fairness and diversity, we prioritized videos containing a wide range of question types and maintained variation in average video duration during selection. Annotators verified the correctness of the QA with video alignment and assigned a quality score (0–5). Results confirm the reliability of our pipeline, achieving 92.1% accuracy and an average score of 4.3.

4. Experiments

In this section, we present our experimental results. We describe our experimental setup in Sec. 4.1, where we introduce the models used and the evaluation setting. The main results are presented in Sec. 4.2. Finally, we further discuss the audio-only evaluation in Sec. 4.3.

4.1. Experimental Setup

Evaluated MLLMs. We evaluate our benchmark on a range of state-of-the-art omni-models that can jointly process audio and video signals. Specifically, the evaluated models include VideoLLaMA2.1-AV [5], video-SALMONN 2+ (7B, 72B) [31], MiniCPM-o 2.6 [41], Qwen2.5-Omni (3B, 7B) [37], Qwen3-Omni-30B (Instruct, Thinking) [38]. These models range in size from 3B to 72B parameters, covering a broad spectrum of model capacities. In addition, we evaluate EgoGPT [40], a model specifically tailored for egocentric video understanding. For fairness, we exclude Gemini [7] from evaluation, since the captions in our dataset were annotated using Gemini 2.5 Flash, which may introduce potential bias.

Evaluation Metrics. Since the QA pairs are in open-ended form, we employ GPT-5 [24] as an automatic validation tool to assess model performance on EgoSound. GPT-5 [24] evaluates the factual consistency between each model's predicted answer and the ground-truth reference (correct answer). Following prior work [5, 36], we define two evaluation metrics for model predictions: Accuracy (0–100%), the percentage of predicted answers judged as "correct"; and Score (0–5), the degree of semantic consistency between the predicted and reference answers, where 5 indicates a fully correct answer and 0 is completely wrong.

Human Evaluation. We recruited two English-proficient evaluators, each holding a bachelor's degree and possessing solid research experience in computer vision, to conduct the human evaluation of our benchmark. Following the protocol in Sec. 3.5, we sampled a subset of 350 QA pairs for evaluation. Additionally, we randomly shuffled the QA pairs to prevent similar questions from appearing consecutively. The human evaluation results are shown in Tab. 3.

4.2. Main Results

We benchmark nine representative MLLMs on the EgoSound, and the quantitative results are summarized in Tab. 3. Overall, our experiments reveal that egocentric sound understanding remains a formidable challenge for current MLLMs, despite their strong progress in vision–language integration.

① **EgoSound poses a significant challenge in current MLLMs.** Human evaluators achieve an average accuracy of 83.9%, whereas the best model, Qwen3-Omni-Thinking-30B [38], reaches only 56.7%, indicating a large gap of over 27 points. This demonstrates that while MLLMs can align vision and language, they still struggle to ground sound cues for reliable perception and reasoning. The performance gap confirms the unique difficulty of multisensory understanding in first-person settings, where sound and vision are deeply entangled.

② **Well-validated ability drops on audio.** Existing MLLMs are well-known for their strong perception ability within the visual domain, such as recognizing ob-

Methods	Characteristics	Counting	Temporal	Location	Identification	Causality	Cross-Modal	Average
Human	95.6 / 4.4	82.5 / 4.0	67.5 / 3.3	81.6 / 3.8	69.7 / 3.6	95.2 /4.3	90.5 / 4.1	83.9 / 3.9
Open-source Models								
VideoLLaMA2.1-AV-7B [5]	15.5 / 1.3	27.7 / 1.4	29.1 / 1.7	21.7 / 1.2	19.4 / 1.1	16.5 / 1.1	13.6 / 0.9	20.5 / 1.3
video-SALMONN 2+ -7B [31]	29.2 / 2.0	40.1 / 2.1	31.5 / 1.8	26.9 / 1.5	34.1 / 1.8	46.9 / 2.5	40.6 / 2.2	36.0 / 2.0
video-SALMONN 2+ -72B [31]	38.4 / 2.5	51.3 / 2.8	37.3 / 2.1	35.4 / 1.9	43.7 / 2.3	63.5 / 3.3	51.1 / 2.8	46.6 / 2.5
Qwen2.5-Omni-3B [37]	28.3 / 1.9	48.7 / 2.5	31.5 / 1.8	26.7 / 1.5	17.7 / 1.0	46.8 / 2.4	33.7 / 1.9	33.9 / 1.9
Qwen2.5-Omni-7B [37]	34.1 / 2.1	**56.3** / 2.9	31.4 / 1.7	30.7 / 1.6	23.4 / 1.3	55.3 / 2.8	42.6 / 2.3	39.8 / 2.1
MiniCPM-o 2.6-8B [41]	33.5 / 2.2	50.7 / 2.7	36.4 / 2.0	26.5 / 1.4	41.6 / 2.2	48.6 / 2.5	42.2 / 2.3	40.4 / 2.2
Qwen3-Omni-Instruct-30B [38]	45.8 / 2.8	55.0 / 2.9	**50.8** / 3.0	40.1 / 2.2	49.8 / 2.6	65.2 / 3.4	51.6 / 2.8	51.9 / 2.8
Qwen3-Omni-Thinking-30B [38]	**55.0** / 3.1	49.0 / 2.6	46.7 / 2.7	**50.0**/ 2.6	**59.5** /3.0	**73.3** / 3.7	**54.5** / 2.9	**56.7** / 3.0
Egocentric Models								
EgoGPT-7B [40]	21.2 / 1.8	52.6 / 2.8	38.7 / 2.2	26.1 / 1.5	26.7 / 1.4	41.9 / 2.2	28.9 / 1.7	34.3 / 2.0

Table 3. **Evaluation results of MLLMs on EgoSound.** The best results are marked in **bold**, and the second-best are underlined. Characteristics, Counting, and Temporal measure the model's ability to perceive the intrinsic properties of sound. Location, Causality, Reasoning, and Cross-Modal go beyond this, further evaluating the model's multimodal perception and reasoning capabilities.

ject characteristics or identifying approximate locations. However, when evaluated on audio-centric tasks, these well-validated abilities show a clear performance degradation. For instance, models including VideoLLaMA2.1-AV-7B [5], Video-SALMONN 2+ [31], Qwen2.5-Omni [37], and MiniCPM-o [41] perform notably worse on tasks involving sound characteristics and spatial localization, often even below their results on counting, causality, or cross-modal reasoning. This reveals that despite supporting audio inputs, current MLLMs still lag significantly in fine-grained auditory perception.

③ **Model scale's impact on performance.** Larger models generally show improved performance, but this scaling advantage does not eliminate all performance gaps. For instance, video-SALMONN 2+ 72B [31] significantly outperforms its 7B counterpart (46.6% vs. 36.0%), and Qwen2.5-Omni-7B [37] exceeds the 3B version by 5.9% percentage points. Similarly, models with a larger scale (e.g., 30B) perform better than those in the smaller scale range (3B/7B/8B). This is likely because larger models are often trained with more diverse data, leading to better generalization ability. However, this trend is not absolute. The performance difference between VideoLLaMA2.1-AV-7B [5] and Qwen2.5-Omni-3B [37], as well as the fact that Video-SALMONN 2+ 72B [31] fails to surpass Qwen3-Omni-Instruct/Thinking-30B [38], clearly indicates that simply increasing model size does not guarantee superior results.

④ **Egocentric pretraining does not help.** The EgoGPT-7B model [40], which is specifically adapted for egocentric data, achieves an average accuracy of 34.3%, substantially lower than the best Qwen3-Omni-Instruct/Thinking-30B [38] models (above 50%) and even inferior to other models of comparable scale. Although this is not an absolutely fair comparison, the relatively poor performance of EgoGPT-7B suggests that pretraining or finetuning on egocentric data alone does not directly address models' limitations in sound-related understanding. This observation further underscores the importance of developing modality-balanced MLLMs that can jointly reason over vision, language, and audio. We also note that a broader conclusion would ideally rely on multiple egocentric models; however, the current availability of open-source egocentric models remains limited. This again highlights the need for more egocentric-oriented MLLMs to advance research in multisensory egocentric understanding.

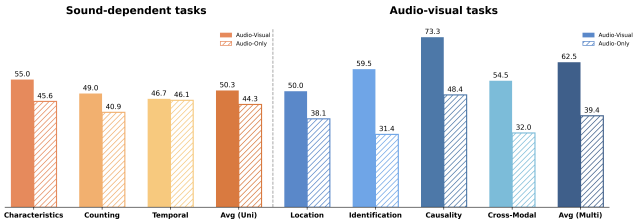

Figure 4. **Accuracy comparison on EgoSound for Qwen3-Omni-Thinking [38] with audio–visual *vs* audio-only input.** Sound-dependent three tasks results are shown in the left (orange), while the right (blue) shows the results for other four tasks that depend on both visual and audio input.

4.3. Audio-Only Evaluation

To investigate the role of visual information in EgoSound, we conduct an audio-only ablation that removes visual inputs while keeping all other settings identical. Specifically, we feed the same questions to the same model, but instead of providing both RGB frames and audio, we retain only the audio track during inference. Following the results in Tab. 3, we select the best-performing model, i.e., Qwen3-Omni-Thinking-30B [38], to perform this ablation. A comparison between dual-modality (audio + visual) and single-modality (audio-only) performance is summarized in Fig. 4.

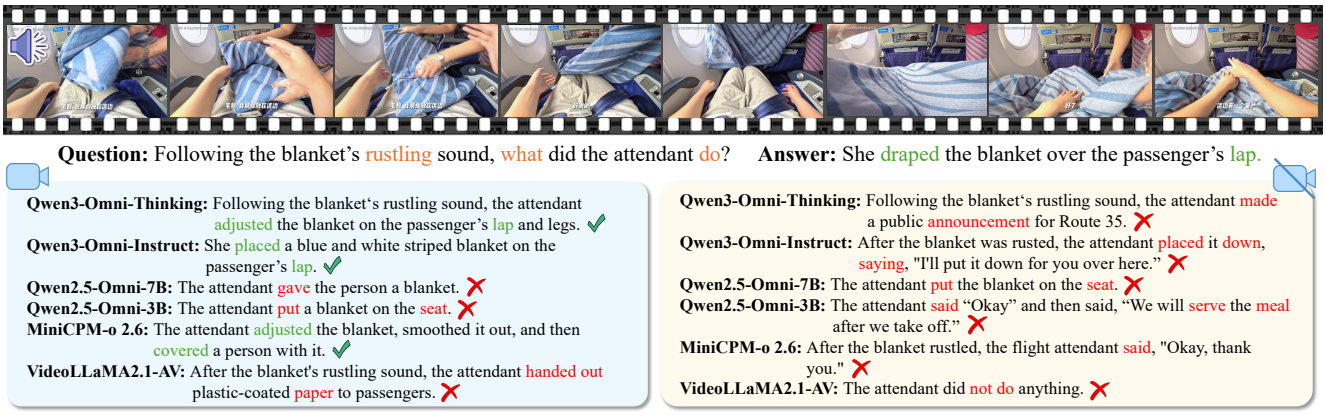

Figure 5. **Comparison of Cross-Modal Reasoning with and without visual input.** The video shows an egocentric airplane scene in which a flight attendant handles a blanket for the passenger. The question asks what happens after the rustling sound produced during this action. The left side presents model outputs with audio-visual frames; the right side presents outputs with audio alone.

From the results, we highlight two main observations: 1) **Sound-dependent tasks remain stable.** For sound-only QA tasks (Characteristics, Counting, and Temporal) , the model maintains relatively stable performance, with only slight drops from 50.3% to 44.3% on average. This is consistent with intuition: when a task depends solely on auditory cues, removing visual inputs should not substantially affect model accuracy. 2) **Audio-visual tasks degrade significantly.** For tasks requiring joint reasoning (Location, Identification, Causality, and Cross-Modal) , we observe a significant decline in performance under the audio-only condition. For example, performance dropped by 28.1 and 24.9 points in Localization and Causality respectively, and the average accuracy also decreased by more than 20%. These findings validate that our audio-visual tasks indeed require integrated reasoning over both modalities, rather than relying on audio alone. Finally, we note that although the model's performance on audio-visual tasks drops noticeably under audio-only input, it does not completely fail. This mirrors the human condition where visually impaired individuals can still interpret the world through auditory cues alone, suggesting that partial reasoning remains possible even without visual modality.

Further analysis of the visual modality. To further illustrate why visual signals is essential for joint reasoning, we visualize representative examples in Fig. 5, comparing model predictions under different input modalities. With dual-modality input (left), the strongest models (Qwen3-Omni-Thinking [38], Qwen3-Omni-Instruct [38], and MiniCPM-o 2.6 [41]) successfully associate the rustling sound with the moment the blanket is being adjusted and correctly ground the subsequent action, identifying that the attendant places or drapes the blanket over the passenger's lap. In contrast, under the audio-only setting (right), all of these models fail to provide the correct answer. This gap clearly demonstrates the importance of cross-modal cooperation for comprehensive first-person scene understanding. Additionally, we observe that weaker models fail under both input conditions, often mislocalizing the action or hallucinating unrelated events. These observations confirm that our QA tasks pose significant challenges to current MLLMs and serve as a realistic and demanding testbed for multisensory egocentric reasoning.

5. Conclusion

In this work, we introduced the EgoSound, the first benchmark designed to systematically evaluate egocentric sound understanding in Multimodal Large Language Models. By unifying the data from Ego4D and EgoBlind and establishing a seven-task taxonomy, spanning sound characteristics, sound event counting, temporal information, spatial location, sound source identification, inferential causality, and cross-modal reasoning, EgoSound offers a comprehensive and realistic testbed for multisensory egocentric intelligence. Through large-scale evaluation across nine state-of-the-art MLLMs, we reveal that while existing models exhibit emergent auditory reasoning abilities, they continue to struggle with fine-grained audio perception and first-person multimodal joint reasoning. Our audio-only ablation studies further highlight the indispensable role of visual cues in many sound-related reasoning scenarios, reinforcing the need for balanced multimodal learning across vision, audio, and language. EgoSound aims to spark deeper exploration into multisensory modeling, especially in the first-person perception domain, for boosting better and robust human-aligned intelligence. We hope this benchmark not only exposes current limitations but also catalyzes progress toward next-generation MLLMs that can truly see, hear, and understand the world from a first-person view.

Acknowledgments

This work is supported by TeleAI.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4

[3] Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. Savvy: Spatial awareness via audio-visual llms through seeing and hearing. *arXiv preprint arXiv:2506.05414*, 2025. 3

[4] 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 *CVPR*, 2024. 2, 3

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3, 6, 7

[6] Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *arXiv preprint arXiv:2506.07016*, 2025. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 3, 5, 6

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2

[9] Chenyou Fan. Egovqa - an egocentric video question answering benchmark dataset. In *ICCV (Workshops)*, 2019. 2, 3

[10] Yuqian Fu, Runze Wang, Yanwei Fu, Danda Pani Paudel, and Luc Van Gool. Cross-view multi-modal segmentation@ ego-exo4d challenges 2025. *arXiv preprint arXiv:2506.05856*, 2025.

[11] Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *ICCV*, 2025.

[12] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. Amego: Active memory from long egocentric videos. In *ECCV*, 2024. 2, 3

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 3, 4, 6

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024. 2

[15] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *CVPR*, 2024. 2

[16] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. *TPAMI*, 2025. 2

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 5

[18] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*, 2022. 2, 3

[19] Kailing Li, Qi'ao Xu, Tianwen Qian, Yuqian Fu, Yang Jiao, and Xiaoling Wang. Clivis: Unleashing cognitive map through linguistic-visual synergy for embodied visual reasoning. *arXiv preprint arXiv:2506.17629*, 2025.

[20] 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. *arXiv preprint arXiv:2508.10729*, 2025. 2, 3

[21] Mohammad Mahdi, Yuqian Fu, Nedko Savov, Jiancheng Pan, Danda Pani Paudel, and Luc Van Gool. Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. *arXiv preprint arXiv:2511.20186*, 2025.

[22] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024. 2

[23] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2023. 1, 2, 3

[24] OpenAI. Gpt-5 system card, 2025. Accessed: 2025-08-10. 6

[25] Jiancheng Pan, Runze Wang, Tianwen Qian, Mohammad Mahdi, Yanwei Fu, Xiangyang Xue, Xiaomeng Huang, Luc Van Gool, Danda Pani Paudel, and Yuqian Fu. V2-sam: Marrying sam2 with multi-prompt experts for cross-view object correspondence. *arXiv preprint arXiv:2511.20886*, 2025. 2

[26] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *CVPR*, 2025. 2

[27] 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 *CVPR*, 2025. 1, 2, 3

[28] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *CVPR*, 2023. 3

[29] Ivan Rodin, Tz-Ying Wu, Kyle Min, Sharath Nittur Sridhar, Antonino Furnari, Subarna Tripathi, and Giovanni Maria Farinella. Easg-bench: Video q&a benchmark with egocentric action scene graphs. *arXiv preprint arXiv:2506.05787*, 2025. 2, 3

[30] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *NeurIPS*, 2023. 3

[31] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Captioning-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025. 2, 3, 6, 7

[32] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *ICCV*, 2025. 3

[33] Weiguo Wang, Andy Nie, Wenrui Zhou, Yi Kai, and Chengchen Hu. Teaching physical awareness to llms through sounds. *arXiv preprint arXiv:2506.08524*, 2025. 3

[34] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, 2023. 2

[35] Zichen Wen, Yiyu Wang, Chenfei Liao, Boxue Yang, Junxian Li, Weifeng Liu, Haocong He, Bolong Feng, Xuyang Liu, Yuanhuiyi Lyu, et al. Ai for service: Proactive assistance with ai glasses. *arXiv preprint arXiv:2510.14359*, 2025. 2

[36] Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind people. *arXiv preprint arXiv:2503.08221*, 2025. 1, 2, 3, 4, 6

[37] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2, 3, 6, 7

[38] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2, 3, 6, 7, 8

[39] Qi'ao Xu, Tianwen Qian, Yuqian Fu, Kailing Li, Yang Jiao, Jiacheng Zhang, Xiaoling Wang, and Liang He. Togbench: Task-oriented spatio-temporal grounding in egocentric videos. *arXiv preprint arXiv:2512.03666*, 2025. 2

[40] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *CVPR*, 2025. 2, 3, 6, 7

[41] 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. *arXiv preprint arXiv:2408.01800*, 2024. 2, 3, 6, 7, 8

[42] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms for video qa. *arXiv preprint arXiv:2410.07177*, 2024. 2

[43] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3

[44] Deheng Zhang, Yuqian Fu, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, et al. Egonight: Towards egocentric vision understanding at night with a challenging benchmark. *arXiv preprint arXiv:2510.06218*, 2025. 2

[45] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*, 2024. 3

[46] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 3

[47] 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 *CVPR*, 2025. 2

[48] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3