

# EgoSound: Benchmarking Sound Understanding in Egocentric Videos

## Supplementary Material

### 1. More Construction Details

This section provides additional details on the prompting strategies used in our multi-stage data curation pipeline. As described in Sec.3.3 of the main paper, EgoSound relies on structured human–interaction annotations, fine-grained audio–visual caption generation, and a robust QA construction process. Below, we present the exact prompt designs that operationalize these stages.

**Audio-Visual Caption Generation Prompt**

You are an expert audio-visual analyst specializing in egocentric videos. You will receive two inputs: (1) Human Interaction Annotations (HIA) and (2) the corresponding video segment. Your task is to generate detailed and unambiguous captions describing all meaningful sound events by combining auditory signals with visually grounded human-object interactions.

**Your Core Execution Logic:**  
Your process **MUST** begin by identifying all meaningful sounds, including speech and environmental sounds. For each sound, you must locate its position in time, link it to relevant visual actions indicated by the HIA (If Have), and then describe the event in detail from different tasks listed in **Your Goal** below. Each description must integrate audio perception, visual grounding, temporal alignment, and causal reasoning.

**Your Goal:** Detailed Sound Analysis  
Focus on all sounds (human speech and environmental sound). In your description, you must address the following seven tasks as much as possible:

- 1. Sound Characteristics:** Describe the sound's intrinsic acoustic properties, including its volume, texture, or timbre. (e.g., "a sharp clink," "a soft rustle," "a loud bang," "a faint whisper.")
- 2. Counting:** Number of distinct sound events or repetitions of a single sound.
- 3. Temporal Attribute:** Occurrence timing, duration, onset/offset characteristics, or volume variation. (e.g., "a continuous sound, and its volume was gradually decreasing over the 5-second period.")
- 4. Spatial Location:** Relative location of the source to the camera (e.g., "knock from the left, behind, 1m" "distant siren from the right, 10m" "nearby footstep from behind, 0.5m")
- 5. Sound Source Identification:** What object or action (visual event) created the sound? (e.g., "a cup hitting the table," "birds chirping," "people talking," "As the hand places the mug on the counter, it makes a loud thud.")
- 6. Inferential Causality:** Why a sound or speech event occurred, focusing on the underlying reason, motivation, or context. This requires leveraging the chronological context (surrounding events) to infer the purpose. (e.g., "The user closed the window to block out the loud shouting man outside.")
- 7. Cross-Modal Reasoning:** How audio events influence visual understanding or how visual context helps explain audio, requiring cross-modal inference.
  - Audio-Guided Visual Reasoning (Sound to Visual): Use audio cues to ground and interpret what is/will happen visually.
  - Visual-Guided Audio Reasoning (Visual to Sound): Use visual cues to ground and interpret what is/will happen in the audio.

**Your Principle:** Filtering and Coherence

- Core Rule: Output must be strictly chronological and non-redundant.
- Focus on: Interactions, human actions, and clear cause-effect audio-visual relationships.
- MUST IGNORE: Completely ignore silent segments, idle scenes (e.g., just standing), or repetitive actions without significant meaning.
- DO NOT: Narrate per-second. Only describe meaningful audio events.
- Context: DO NOT hallucinate visual or audio details.
- Termination: MUST stop generating once you have described all meaningful sound events.

**Your Final Output:** Chronological Exceedingly Detailed Audiovisual Caption  
Your final output must be exceedingly detailed and form a coherent caption that combines audio and visual reasoning.

- Structure: Use the chronological timeline format: [start time – end time] event description. Use seconds as timestamps (e.g., 00:12 - 00:15).
- Format: Each line MUST represent one distinct, non-overlapping audio event.
- Speech Handling: If an event includes full speech/dialogue, the transcribed text MUST be clearly embedded in the description using quotation marks or a clear label (e.g., "The user says: ...").
- Style: Be concise but highly informative. Avoid redundancy and trivial details like "standing," "breathing," "background wall," or "nothing happens."

Figure 1. Prompt of Audio-Visual Caption Generation.

#### 1.1. Prompt for Caption Generation

To transform annotated human–object and human–human interactions into detailed, sound-centric descriptions, we

follow [3] and design a specialized prompt that instructs the model to generate chronological audio–visual captions grounded in both audio cues and visual context. Fig. 1 shows the full prompt template used for audio–visual caption generation.

#### 1.2. Prompt for QA Pairs Generation

Building on the generated captions, we design a new prompt to construct sound-centric QA pairs. The prompt instructs the model to produce questions spanning the seven core tasks defined in EgoSound, ensuring that the resulting QA pairs cover intrinsic audio perception, spatial reasoning, causal inference, and cross-modal understanding. Fig. 2 presents the complete prompt used for QA construction.

### 2. More Analysis and Visualization Results

In this section, we provide additional analyses and visualizations to complement the results reported in the main paper. Beyond quantitative comparisons, these materials offer deeper insight into how current MLLMs behave across different sound-centric tasks in EgoSound. We include the implementation details of our main experiments, the full prompt used for OpenQA evaluation, and representative QA examples from both Ego4D and EgoBlind to illustrate the diversity and complexity of our data sources. We also present a range of failure cases that reveal systematic weaknesses in temporal grounding, spatial localization, sound source identification, cross-modal alignment, and acoustic perception. Together, these analyses provide a more comprehensive understanding of the challenges posed by egocentric sound reasoning and highlight key opportunities for improving future multimodal models.

#### 2.1. Implement Details of Experiments

For fair and reproducible evaluation, we execute different model sizes on hardware appropriate to their computational requirements. Specifically, all small and medium-sized models—including VideoLLaMA2.1-AV-7B [1], video-SALMONN 2+-7B [5], Qwen2.5-Omni-3B [7], Qwen2.5-Omni-7B [7], and MiniCPM-o 2.6-8B [9]—are run on NVIDIA RTX 4090 GPUs. Larger models—including video-SALMONN 2+-72B [5], Qwen3-Omni-Instruct-30B [8], and Qwen3-Omni-Thinking-30B [8]—are executed on NVIDIA H200 GPUs to ensure stable inference and prevent memory bottlenecks. All models are evaluated under the same zero-shot, single-round inference protocol. For visual inputs, video frames are sampled at dataset-specific rates between 0.5–1 fps, and no maximum frame

### Q&A Generation Prompt

You are an expert AI for sound-centric information extraction. Read the detailed audiovisual caption and generate high-quality Question-Answer (QA) pairs that cover the seven core sound-centric tasks. Base every question and answer solely on information contained in the caption or on inferences that naturally follow from its context. Use the accompanying video frames for double validation: confirm that each QA pair is fully supported by visual evidence, correcting or discarding any detail that cannot be verified. Do not introduce content beyond what is grounded in either the caption or the video frames, and ensure all QA pairs remain factual, precise, and hallucination-free.

#### Core Objective:

Every question you generate must target one of the following seven core tasks, as defined below:

#### SEVEN Tasks (Question Focus):

**1. Sound Characteristics:** Ask about the sound's acoustic qualities, such as its volume or texture. (e.g., "What is the intensity and volume of distant traffic noise?", "Was the 'tap' at 01:10 sharp, muffled, dull, or resonant?")

**2. Counting:** Ask about the count, number of repetitions, or frequency of sound events. (e.g., "How many types of environmental sounds (excluding speech) were simultaneously present at the 00:45 mark?", "How many times did the 'clink' sound occur?", "How many distinct, quick footsteps were recorded between 02:00 and 02:05?")

**3. Temporal Attribute:** Ask about the sound's timing, duration, or volume fluctuation. (e.g., "When did the 'clink' sound start and how long did it last?", "What is the duration and volume change of the high-frequency whining sound heard between 01:30 and 01:35?")

**4. Spatial Location:** Ask about the sound's precise location (Direction & Distance) relative to the camera. (e.g., "What was the spatial location (direction and estimated distance in meters) of the faint rustling sound at 00:15?", "Did the close-range speech originate from the left side, the right side, or directly in front of the camera?")

**5. Sound Source Identification:** Ask what specific object, person, or action generated the sound. (e.g., "What object generated the 'clink' sound at 00:15?", "What was the source of the high-pitched metallic scraping sound at 00:15?")

**6. Inferential Causality:** Ask why a sound or speech event occurred, focusing on the underlying reason, motivation, or context. This requires leveraging the chronological context (surrounding events) to infer the purpose. (e.g., "What was the likely reason for the car's sudden horn honk at 01:10?", "Why did the speaker laugh at 00:45?", "Based on the preceding events, why did the mom say 'I'm so proud?'")

**7. Cross-Modal Reasoning:** Ask how audio events influence visual understanding or how visual context helps explain audio, requiring cross-modal inference.

- **Audio-Guided Visual Reasoning (Sound to Visual):** Use audio cues to ground and interpret what is/will happen visually. (e.g., "After the loud crash sound at 00:22, what object most likely fell?", "When the barking sound occurs, where is the dog located in the scene?")

- **Visual-Guided Audio Reasoning (Visual to Sound):** Use visual cues to ground and interpret what is/will happen in the audio. (e.g., "When the man slammed the door at 00:45, what sound followed?", "After the woman claps her hands, what sound follows and what does it suggest about the environment?")

#### Execution Principles:

- **Quality First:** Emphasis is placed on QA quality, not on absolute quantity. Only generate meaningful questions that have a clear answer within the text.  
- **Strictly Grounded:** All QA pairs must be inferred from the provided caption text. DO NOT hallucinate or invent details not present in the text.  
- **Filter Irrelevance:** If a line in the caption does not contain specific details related to the 7 tasks (e.g., it only says "The user is walking" with no sound description), skip that line and do not generate a QA for it.

- **No Duplication:** Avoid generating multiple, repetitive questions about the exact same event detail.  
- **Validate and Prune:** After generating the QA pairs, re-check each question and its answer against the caption and video frames; if a QA pair contains any incorrect, unsupported, or ambiguous information, discard it immediately.

#### Output Format (JSON):

You MUST return your answer as a single JSON list. Each QA pair must be an object within this list.

#### JSON Format Example:

```
{  
  "timestamp": "00:55 - 01:00",  
  "context": "[00:55] A man outside the window shouts loudly. [00:58] The user closes the window with a definitive thud.",  
  "question_type": "Inferential Causality",  
  "question": "Based on the preceding event, what was the inferred reason for the user closing the window at 00:58?",  
  "answer": "The user closed the window to block out the loud shouting man outside."  
},  
  "timestamp": "00:20 - 00:25",  
  "context": "[00:20] The user's hand is seen dropping a key fob. [00:21] A loud, metallic clang is heard from the floor, 0.5m directly below the camera.",  
  "question_type": "Sound Characteristics",  
  "question": "What was the quality and volume of the sound made by the dropped key fob at 00:21",  
  "answer": "The sound was a 'loud, metallic clang!'"  
}
```

Figure 2. Prompt of QA Pairs Generation.

limit is enforced, preserving the temporal continuity of each egocentric sequence. Audio streams are fed at their original sampling rate and are not modified. Each model receives identical multimodal inputs consisting of the sampled video frames, synchronized audio, and the question text.

### OpenQA LLM-as-Judge Prompt

"role": "system",  
"content": "You are an intelligent chatbot designed for evaluating the correctness of generative outputs for video-based question-answer pairs.  
Your task is to compare the predicted answer with the correct answer and determine if the predicted answer is correct or not. Here's how you can accomplish the task:  
-----

#### ##INSTRUCTIONS:

- Focus on the correctness and accuracy between the predicted answer and the correct answer. The predicted answer should not contain any misinterpretations or misinformation.
- The predicted answer must be factually accurate and align with the video content.
- The predicted answer and the correct answer may differ in language. Make sure to translate them into the same language and compare their semantic consistency.
- The predicted answer can be considered a valid answer to the question.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

"role": "user",

"content": "Please evaluate the following video-based question-answer pair:

Question: {question}

Correct Answer: {answer}

Predicted Answer: {pred}

Provide your evaluation only as a correct/incorrect prediction along with the score where the score is an integer value between 0 (fully wrong) and 5 (fully correct). The middle score provides the percentage of correctness.

Please generate the response in the form of a Python dictionary string with keys 'binary\_pred' and 'score', where value of 'binary\_pred' is a string of 'correct' or 'incorrect' and value of 'score' is in INTEGER, not STRING.

For example, your response should look like this: {'binary\_pred': 'correct', 'score': 4}."

Figure 3. Prompt of LLM-as-Judge. The prompt takes as input the question, the correct answer (answer), the model's prediction (pred), to produce the resulting evaluation.

## 2.2. Prompt for OpenQA Evaluation

Given the subjective nature of open-ended responses, we adopt GPT-5 [4] as an automated judge to provide consistent and scalable evaluation. The LLM judge assesses the factual consistency of each model prediction relative to the ground truth, following the prompt design illustrated in Fig. 3.

## 2.3. QA examples Visualization

Fig. 4 and Fig. 5 present representative QA examples from the Ego4D [2] and EgoBlind [6] subsets of EgoSound. Each QA pair is constructed from both the detailed clip caption shown in the rightmost column and the corresponding video frames, ensuring that every question and answer is grounded in the underlying audiovisual evidence. The examples follow the seven egocentric sound tasks defined in the main paper: the intrinsic sound properties (Sound Characteristics, Counting, Temporal Attribute) and the multi-modal perception and reasoning tasks (Spatial Location, Sound Source Identification, Inferential Causality, Cross-Modal Reasoning). These visualizations illustrate how the caption-to-QA pipeline faithfully anchors sound events to their temporally aligned egocentric audio-visual context.

## 2.4. Failure Cases Analysis

To further understand the limitations of current MLLMs on EgoSound, we provide a detailed analysis of representative failure cases across the seven sound-centric tasks. The following examples highlight common error patterns—such as misaligned temporal grounding, incorrect spatial reason-



Characteristics	Sound Characteristics	Q: How are the chip sounds characterized while the first player explains the rules? A: They are subtle and intermittent clinking noises.	[00:05 - 00:08] The first player explains tournament rules while sorting red and yellow chips on the table, generating subtle, intermittent clinking noises.
		Q: What was the sound quality when the card was dropped on the glass table? A: A soft thud.	[00:53 - 00:54] A soft thud is heard as the second player drops a playing card onto the glass table.
Counting	Sound Event Counting	Q: How many distinct clinks were produced when the first player placed the two red chips? A: Two clinks—one for each chip placement.	[00:31 - 00:32] As the first player says, "Make sure the stacks look clean," he precisely places two red chips on an existing stack, each placement creating a distinct, soft clink.
		Q: How many clinks followed the statement "I'm going to raise by two," and what did they correspond to? A: Two clinks, corresponding to two yellow chips being placed on the table.	[02:15 - 02:16] The second player announces, "I'm going to raise by two," directly followed by two distinct clinking sounds as he places two yellow chips onto the table.
Temporal	Temporal Information	Q: What was the sequence of audible actions between 01:21 and 01:25? A: A slight rustle/shuffle as the hand moved, followed by the sound of a single card being placed face down.	[01:21 - 01:25] The first player talks about Uno; a slight shuffling or rustling is heard as the camera-wearer moves their hand, then a playing card is placed face down onto the table.
Location	Spatial Location (Direction & Distance)	Q: From which direction relative to the camera did the rapid card shuffling originate? A: From the left side, where the camera-wearer's hands entered the frame.	[00:19 - 00:22] The camera-wearer's hands enter the frame from the left and perform a series of quick, audible card shuffles, followed by the distinct sound of cards being placed firmly onto the glass table.
Identification	Sound Source Identification	Q: What generated the faint clinking sound at the start? A: The first player dropping a red Connect 4 chip onto the glass coffee table.	[00:00 - 00:03] First player says, "is always the same, it's only one chip." Second player replies "Ah," first player says "Yeah." Simultaneously, the first player manipulates red and yellow Connect 4 chips on the glass coffee table, producing faint, distinct clinking sounds as he drops a red chip.
Causality	Inferential Causality	Q: Why did the second player say, "Wow, didn't even get very far"? A: Because the first player folded quickly, ending the hand early.	[00:53 - 00:54] The second player reacts, "Wow, didn't even get very far," immediately followed by a soft thud as he drops a playing card onto the glass table.
		Q: Why did the first player place a yellow chip on the table at this moment? A: To award the second player the chip he had just won.	[01:00 - 01:01] The first player says, "See, you won your chip." He then picks up a yellow chip and places it on the table, causing a light clinking sound.
Cross-Modal	Sound to Visual Visual to Sound	Q: When the first player placed a red chip on a stack during this sequence, what sound followed? A: A distinct clink, followed by additional soft clinks as he moved other chips.	[00:47 - 00:50] The first player says, "I am going to," followed by a distinct clink as he places a red chip on a stack. He adds, "And uh," with more soft clinks as he moves chips, concluding with, "I'm gonna fold."
		Q: How did the chip placements at 02:27 and 02:29 illustrate the betting explanation being given? A: By adding two chips and then another chip as the rule was described, the second player visually demonstrated adding chips on top of the matching bet to represent a raise.	[02:21 - 02:29] While the first player explains how raising works, the second player places two yellow chips at 02:27 and another yellow chip at 02:29, each producing a clear clink and visually demonstrating the "raise" action.

Figure 4. A visualization of representative QA examples, the video source is from the EGO4D [2] dataset.

ing, audio–visual mismatches, off-screen sound misidentification, and difficulties in perceiving low-quality or rapid speech.

**Failure case of Sound Characteristics.** Fig. 6 highlights a failure case in sound characteristics perception, where the voice from the phone is rapid, synthesized, and unclear. The models exhibit mixed results in handling this distorted audio. Some models detect the poor acoustic quality, describing the voice as distorted, muffled, and difficult to understand due to static or ambient noise, suggesting issues with the signal or the speaker. Others, however, interpret the voice as clear or loud, indicating good acoustic quality, despite the presence of distortions and background noise.

This case underscores the challenges faced by current MLLMs in accurately perceiving audio with poor quality. In everyday settings, audio quality is often inconsistent,

with varying degrees of distortion and noise. The variety of responses shown in the outputs reveals that current MLLMs struggle to consistently interpret such challenging audio conditions. Therefore, EgoSound, which includes a range of real-life scenarios with varying audio quality, presents a significant challenge to the audio perception capabilities of current MLLMs.

**Failure case of Counting.** Fig. 7 illustrates a case where models struggle with detecting the counting of the phrase “Come here” in an egocentric video segment. The correct answer is that the woman says “Come here” three times in total, but most models fail to capture this accurately. While some models correctly identify two occurrences, others report only one or even zero. The main reason for this discrepancy is that the first instance of “Come here” in the second frame is spoken too quickly and lightly for the models to



Characteristics	Sound Characteristics	Q: What are the volume and mood characteristics of the background music at the start? A: It is loud and cheerful.	[00:00 - 00:05] The camera moves along a city street with the loud, cheerful sound of background music.
		Q: What is the volume quality of the woman's directions? A: Clear conversational volume.	[00:05 - 00:18] The woman provides directions at a clear conversational volume.
		Q: How audible are the sounds of the user's fingers interacting with the phone? A: Faintly audible.	[00:53 - 00:57] The user takes out their smartphone; faint sounds of fingers interacting with the phone are audible. They say, '稍等啊' ('Wait a moment').
Temporal	Temporal Information	Q: What spoken request follows after the user says '那给你'? A: The employee asks for the shipping address: '你把地址给我'.	[00:47 - 00:53] The user says '那给你' ('Then here you go') while handing a brown paper bag; the employee asks, '你把地址给我' ('Give me the address').
Location	Spatial Location (Direction & Distance)	Q: What is the approximate distance of the woman who is giving directions? A: About 2 meters away.	[00:05 - 00:18] A woman, standing about 2 meters away, gives directions at a clear conversational volume: '下来...走这边吗...这块啊'.
		Q: How far from the camera is the employee when speaking at the entrance? A: About 1.5 meters away.	[00:24 - 00:31] At the entrance, an employee inside, about 1.5 meters away, asks, '取快递啊, 寄快递啊?'
Identification	Sound Source Identification	Q: Who provides the final directions near the entrance? A: Another person off-camera.	[00:18 - 00:24] The user walks toward the 'Gainiao Post' entrance; another person off-camera provides final directions. The user replies, '好的谢谢啊...好好好'.
Causality	Inferential Causality	Q: Why did the user say '谢谢嗯' in this segment? A: To thank the woman after she provided directions indicating the destination was just ahead.	[00:05 - 00:18] The user asks where to go, is told the courier station is just ahead, and says '谢谢嗯' ('Thanks, mm-hmm').
		Q: Why does the user request additional layers of bubble wrap? A: Because the item is a fragile, handmade glass cup and they fear it might break.	[00:32 - 00:47] The user explains the item is a fragile, handmade glass cup and requests extra bubble wrap: '能不能就多多给我垫几层那个泡沫纸, 我怕它碎掉'.
		Q: Why does the user say '稍等啊' at this point? A: Because they are searching for the shipping address on their smartphone.	[00:53 - 00:57] The user says '稍等啊' ('Wait a moment') while looking for the address on their phone.
Cross-Modal	Sound to Visual Visual to Sound	Q: As the woman speaks to give directions, what visual action accompanies her speech? A: She points the way forward.	[00:05 - 00:18] While speaking, the woman points the way forward.
		Q: When the user says '那给你', what concurrent action occurs? A: They hand a brown paper bag containing the item to the employee.	[00:47 - 00:53] As the user says '那给你', they hand a brown paper bag to the employee.

Figure 5. A visualization of representative QA examples, the video source is from the EgoBlind [6] dataset.

detect, leading to its omission in their results.

This example highlights the challenges faced by current MLLMs in detecting fast-spoken or unclear audio, even in relatively simple scenarios. The inability to capture rapid speech accurately, particularly in noisy or dynamic environments, poses significant obstacles for the audio perception capabilities of these models.

**Failure case of Temporal Attribute.** Fig. 8 illustrates a failure case in the temporal attribute task, where models struggle to accurately identify the duration of the sound event. The correct duration of the audible description is from 00:09 to 00:14, lasting approximately 5 seconds. However, models fail to align the audible description temporally, often outputting timestamps that extend beyond the actual video length. For example, some models incorrectly predict durations as long as 23 seconds or more, even though the entire video is only 16 seconds long. This error

indicates that the MLLMs fail to accurately align the auditory event with its corresponding time frame within the video. The discrepancy reveals a critical challenge in the Temporal Attribute task: models' inability to precisely capture the timing and duration of sound events, even when the total video duration is known.

This case highlights the need for better temporal grounding and understanding in multimodal models. Accurate identification of the specific timing of audio cues is crucial for tasks that rely on precise temporal reasoning.

**Failure case of Spatial Location.** Fig. 9 illustrates a failure case in the spatial location task. The correct answer is that the man's corrective explanation originates from roughly two meters directly in front of the camera. However, the models produce highly inconsistent spatial predictions, placing the sound source to the right, to the left, or even behind the camera. A key reason for these errors is



**Question:** How would you describe the acoustic qualities of the voice heard from the phone during the call?

**Answer:** It was rapid and synthesized, with a female timbre, reading at high speed.

**Qwen3-Omni-Thinking:** The voice heard from the phone during the call has very poor acoustic quality. It is heavily distorted, crackling, and filled with static, making it difficult to understand clearly. The sound is muffled and tinny, with a high-pitched, warbling quality that suggests a weak signal or a low-quality speaker.

**Qwen3-Omni-Instruct:** The voice heard from the phone is not the driver's voice, but rather the ambient environmental sounds of the car, such as the engine and road noise. This indicates that the phone is connected to the car's microphone but the driver is not speaking into it.

**MiniCPM-o 2.6:** The voice heard from the phone during the call is clear and distinct, indicating good acoustic quality.

**Qwen2.5-Omni-7B:** The voice from the phone during the call was quiet, but you could still hear the ambient noise around. It was like the person on the other end was in a noisy environment. So, it was a bit hard to hear clearly.

**Qwen2.5-Omni-3B:** The voice heard from the phone during the call is described as having a clear and distinct quality. It's like you can hear the environment around you, but the person on the other end isn't speaking. It's a bit of a mystery, but it's definitely not a normal conversation.

**video-SALMONN 2+ -72B:** The voice is muffled, distorted, and difficult to understand. It sounds like it's coming through a speakerphone or a bad connection.

**video-SALMONN 2+ -7B:** The voice is clear and audible, indicating good audio quality.

**VideoLLaMA2.1-AV:** The voice is loud and clear.

Figure 6. MLLMs fail to consistently perceive low-quality and distorted sound characteristics.



**Question:** How many times did the woman say "Come here" in this segment?

**Answer:** Three times.

**Qwen3-Omni-Thinking:** The woman said "Come here" 2 times in this segment.

**Qwen3-Omni-Instruct:** 1.

**MiniCPM-o 2.6:** The woman said "Come here" twice in this segment.

**Qwen2.5-Omni-7B:** The woman said "Come here" twice in this segment.

**Qwen2.5-Omni-3B:** She said "Come here" twice in this segment.

**video-SALMONN 2+ -72B:** The woman said "Come here" twice.

**video-SALMONN 2+ -7B:** The woman said "Come here" twice.

**VideoLLaMA2.1-AV:** Zero.

Figure 7. MLLMs fail to enumerate the repetitions of a specific phrase in fast-spoken or audio-unclear egocentric videos.

that both the camera and the man are moving during the interaction, causing the spatial cues in the audio to shift dynamically across the segment. Current MLLMs are not yet robust to such egocentric motion: they struggle to integrate evolving visual context with continuously changing auditory directionality. As a result, the models fail to es-



**Question:** What is the duration of the AI's audible description segment?

**Answer:** About 5 seconds, from 00:09 to 00:14.

**Qwen3-Omni-Thinking:** The AI's audible description segment lasts for approximately 23 seconds, running from 00:05 to 00:28.

**Qwen3-Omni-Instruct:** The AI's audible description segment lasts for 9 seconds, from 00:07 to 00:16.

**MiniCPM-o 2.6:** The AI's audible description segment lasts for 12 seconds.

**Qwen2.5-Omni-7B:** The AI's audible description segment lasts from 18.96 seconds to 34.62 seconds.

**Qwen2.5-Omni-3B:** The AI's audible description segment lasts from 18.42 seconds to 25.52 seconds.

**video-SALMONN 2+ -72B:** The AI's audible description segment lasts from 00:00:28 to 00:00:41.

**video-SALMONN 2+ -7B:** The AI's description segment lasts for 1 minute and 20 seconds.

**VideoLLaMA2.1-AV:** The video shows a person taking a picture of themselves in a mall. The person is walking down a hallway with a camera. The AI's audible description segment lasts for 8 seconds.

Figure 8. MLLMs fail to accurately identify temporal boundaries of sound events in short egocentric videos.



**Question:** Where did the man's corrective explanation originate relative to the camera?

**Answer:** From about 2 meters directly in front of the camera.

**Qwen3-Omni-Thinking:** The man's corrective explanation originated from his position to the right of the camera.

**Qwen3-Omni-Instruct:** To the right of the camera.

**MiniCPM-o 2.6:** The man's corrective explanation originated to the right of the camera.

**Qwen2.5-Omni-7B:** The man's corrective explanation originated from the left side of the camera.

**Qwen2.5-Omni-3B:** The man's corrective explanation came from behind the camera.

**video-SALMONN 2+ -72B:** The man's corrective explanation originated from the left side of the frame, as he was standing next to the woman at the claw machine.

**video-SALMONN 2+ -7B:** The man's corrective explanation originated from his right side relative to the camera.

**VideoLLaMA2.1-AV:** The video shows a person holding a basket of toys in front of a machine. The person is seen pressing buttons on the machine.

Figure 9. MLLMs fail to localize sound sources under dynamic egocentric motion.

establish a stable spatial reference frame and cannot reliably infer where the sound originates.

This case highlights a core limitation revealed by EgoSound: accurate sound localization in first-person videos requires models to handle dynamic viewpoint changes, a capability that remains underdeveloped in existing MLLMs.

**Failure case of Sound Source Identification.** Fig. 10 shows a failure case in the Sound Source Identification task. The correct answer is that the pleasant, melodic chime orig-



**Question:** What was the source of the pleasant, melodic chime at 00:48-00:51?

**Answer:** The store's speaker system.

**Qwen3-Omni-Thinking:** The pleasant, melodic chime heard between 00:48 and 00:51 is the sound from the customer's smartphone.

**Qwen3-Omni-Instruct:** The pleasant, melodic chime at 00:48-00:51 was the sound of a bicycle bell.

**MiniCPM-o 2.6:** The pleasant, melodic chime was likely from the cash register.

**Qwen2.5-Omni-7B:** It was from a bell. It's a nice sound.

**Qwen2.5-Omni-3B:** The source of the pleasant, melodic chime at 00:48-00:51 was the door opening.

**video-SALMONN 2+ -72B:** The chime was likely triggered by the Cashier scanning an item.

**video-SALMONN 2+ -7B:** The chime was likely from the cash register as it is common for cash registers to emit such sounds when transactions are completed.

**VideoLLaMA2.1-AV:** The source of the pleasant, melodic chime at 00:48-00:51 is the store's door chime.

Figure 10. MLLMs fail to identify off-screen sound sources without visual grounding.

inates from the store's speaker system, which is not visible in the video frames. However, most models incorrectly attribute the sound to visually present objects, such as the customer's smartphone, a bicycle bell, the door opening, or the cash register. A central reason for these errors is that the actual sound source is off-screen, coming from the store's automated speaker system. Correct identification requires using broader video context—such as noticing that the camera wearer has already moved away from the counter—but many models instead rely only on objects currently visible, leading to systematic misattribution.

This case highlights that MLLMs often struggle when the auditory source is not visually grounded within the current frame, revealing their difficulty in performing sound-context integration and in reasoning beyond immediate visual cues.

**Failure case of Inferential Causality.** Fig. 11 reveals a typical failure pattern in current MLLMs. In this scenario, the female exclaims, "Ah, it fell!" because a hand briefly touches the white cane leaning in the corner, after which the cane slides and eventually falls. Although this causal interaction is clearly visible in the egocentric video, none of the evaluated models identified it. Several models instead focus on the nearby white dog and mistakenly attributed the claim to the dog's actions, while others produced hallucinated explanations entirely unrelated to the scene. Most models did not even recognize that the cane had fallen.

This case shows that, without explicit signals indicating human-object interactions, models often overlook the subtle yet decisive physical events that give rise to spe-



**Question:** Why did the female exclaim, 'Ah, it fell'?

**Answer:** She was reacting to the white cane falling over.

**Qwen3-Omni-Thinking:** Because the white cane she was holding was knocked over by the dog.

**Qwen3-Omni-Instruct:** Because the white dog, Yuki, accidentally knocked over the umbrella that the person was holding.

**MiniCPM-o 2.6:** The female exclaimed, "Ah, it fell" because the dog had fallen down.

**Qwen2.5-Omni-7B:** It seems like the dog might have knocked something over. Maybe it was playing around or just got too excited.

**Qwen2.5-Omni-3B:** It seems like the person was using a cane to help them walk, and the dog got in the way. Maybe the dog bumped into the cane, causing it to fall.

**video-SALMONN 2+ -72B:** Because the white dog bumped into her hand while she was holding the black cane.

**video-SALMONN 2+ -7B:** Because she accidentally dropped the white cane while trying to open the door.

**VideoLLaMA2.1-AV:** Because she accidentally dropped the door handle while she was holding it.

Figure 11. MLLMs fail to infer the causality in egocentric video without prior human interaction information.



**Question:** During the upward pan to the store ceiling as he enters, what does the user say?

**Answer:** He says he has an appointment for 11:15.

**Qwen3-Omni-Thinking:** As the camera pans up to the store ceiling, the person says: "热心人的帮助和自己的努力，终于找到了。感谢路上的热心人" (With the help of kind people and my own efforts, I finally found it. Thank you to the kind people on the way.)

**Qwen3-Omni-Instruct:** He says, "I can find it. As long as I have the address, I can find it. If I can't find it, I'll just ask, right?"

**MiniCPM-o 2.6:** During the upward pan to the store ceiling as he enters, the user says: "热心人的帮助和自己的努力，终于找到了。感谢路上的热心人" (The help from kind people and my own efforts finally led us here. Thank you for the helpful people on the way.)

**Qwen2.5-Omni-7B:** The user says, "欢迎大家评论区留言。" which means "Welcome everyone to leave comments in the comment section."

**Qwen2.5-Omni-3B:** The user says, '欢迎大家评论区留言。' which translates to 'Welcome everyone to leave comments in the comment section.'

**video-SALMONN 2+ -72B:** The user says: 'With the help of kind people and my own efforts, I finally found it. Thank you for your kindness on the road.'

**video-SALMONN 2+ -7B:** The user says: '人有出口口诀，一杖，二问，三鼻耳。个人觉得还是有一'.

**VideoLLaMA2.1-AV:** As he walks up to the store ceiling, he says something about the height of the building. The camera then pans to show the outside of the building, which is made of glass and has a tall structure.

Figure 12. MLLMs fail to temporally align audio with visual cues for cross-modal reasoning.

cific auditory outcomes. This example therefore highlights the necessity of incorporating interaction-aware annotations into the data curation pipeline: by providing precise temporal grounding of human-object and human-human in-

teraction, such annotations supply the contextual structure needed for models to align visual actions with their corresponding sounds and perform more reliable cross-modal causal reasoning.

**Failure case of Cross-Modal Reasoning.** Fig. 12 presents a failure case in the Cross-Modal Reasoning task. The correct answer is that the user says he has an appointment for 11:15 as the camera pans upward to the store ceiling during entry. However, most models produce entirely unrelated responses—many simply repeat the Chinese subtitles overlaid in the video or hallucinate arbitrary dialogue.

The main cause of these errors is that the models fail to temporally align the visual cue (“as he enters, during the upward pan to the ceiling”) with the corresponding audio segment. Instead of grounding the spoken utterance to the correct visual moment, the models rely on visible on-screen text or generic context, leading to cross-modal mismatches. This case highlights that current MLLMs still struggle to integrate visual timing cues with audio content, revealing a core limitation in fine-grained cross-modal reasoning.

## References

- [1] 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. 1
- [2] 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. 2, 3
- [3] Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025. 1
- [4] OpenAI. Gpt-5 system card, 2025. Accessed: 2025-08-10. 2
- [5] 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. 1
- [6] 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. 2, 4
- [7] 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. 1
- [8] 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. 1
- [9] 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. 1