EgoAVU: Egocentric Audio-Visual Understanding

Ashish Seth[1,2], Xinhao Mei[1], Changsheng Zhao[1], Varun Nagaraja[1], Ernie Chang[1], Gregory P. Meyer[1], Gael Le Lan[1], Yunyang Xiong[1], Vikas Chandra[1], Yangyang Shi[1], Dinesh Manocha[2], Zhipeng Cai[1]
[1]Meta, [2]University of Maryland, College Park

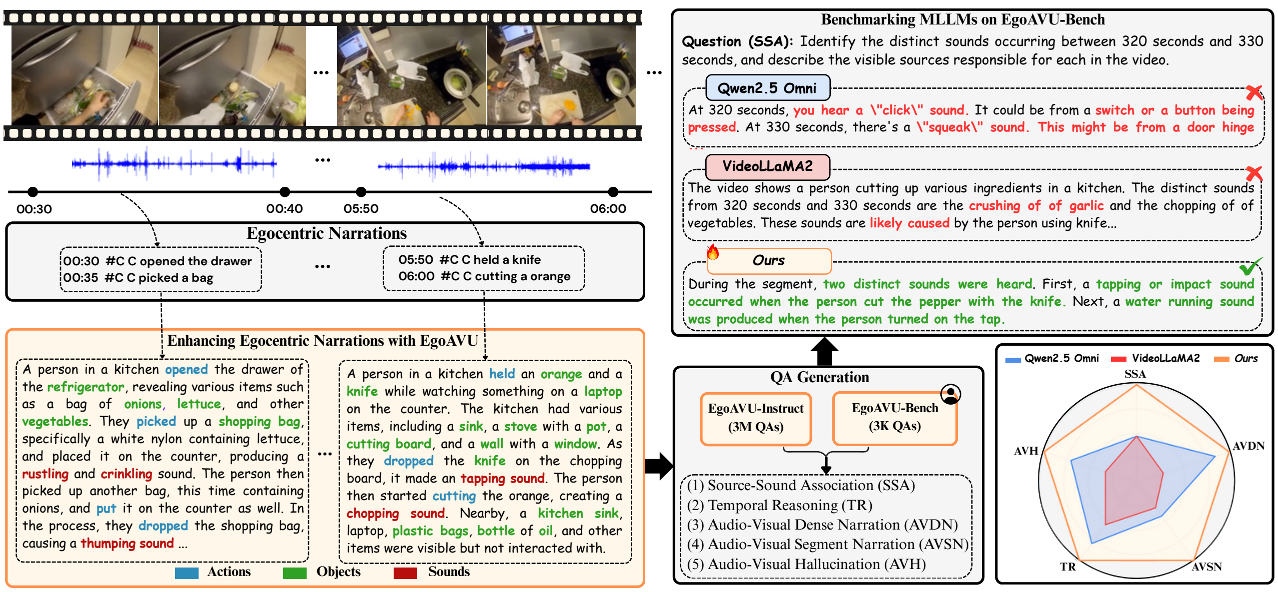

Figure 1. **Overview of EgoAVU.** We introduce EgoAVU, a scalable and automated data engine to enable egocentric audio–visual understanding. EgoAVU enriches existing egocentric narrations by integrating human actions with environmental context, explicitly linking visible objects and the sounds produced during interactions or surroundings. Leveraging this pipeline, we construct *EgoAVU-Instruct (3M QAs)* and *EgoAVU-Bench (3K verified QAs)*, enabling systematic training and evaluation of MLLMs. Models finetuned with *EgoAVU-Instruct* exhibit high audio-visual grounding in egocentric settings.

Abstract

Understanding egocentric videos plays a vital role for embodied intelligence. Recent multi-modal large language models (MLLMs) can accept both visual and audio inputs. However, due to the challenge of obtaining text labels with coherent joint-modality information, whether MLLMs can jointly understand both modalities in egocentric videos remains under-explored. To address this problem, we introduce EgoAVU, a scalable data engine to automatically generate egocentric audio-visual narrations, questions and answers. EgoAVU enriches human narrations with multimodal context and generates audio-visual narrations through cross-modal correlation modeling. Token-based video filtering and modular, graph based curation ensure both the data diversity and quality. Leveraging EgoAVU, we construct EgoAVU-Instruct — a large scale training dataset of 3M samples, and EgoAVU-Bench — a manually verified evaluation split of 3K samples. EgoAVU-Bench clearly reveals the limitation of existing MLLMs: they bias heavily towards visual signals, often neglecting audio cues or failing to correspond audio with the visual source. Finetuning MLLMs on EgoAVU-Instruct effectively solves this issue, enabling up to 113% performance improvement on EgoAVU-Bench. Such benefit can also transfer to other benchmarks such as EgoTempo and EgoIllusion, achieving up to 28% relative performance gain. Project page: https://cs20s030.github.io/EgoAVU/

1. Introduction

Egocentric videos capture rich and dynamic first person audio–visual information centered on daily human activities, such as cooking, painting, or assembling objects [15, 16, 28]. Understanding such data plays a vital role in embodied intelligence and mixed-realities [26, 27, 32, 33].

The highly dynamic camera motion and the limited field of view make comprehensive egocentric video understanding challenging with solely visual cues [7]. This motivates the use of audio information, which provides persistent contextual signals tied to ongoing events. Though recent multimodal language models (MLLMs) can accept both audio and visual inputs [17, 29, 36, 38, 39], *whether they understand the joint dynamic of audio–visual signals in egocentric videos* remains an open question.

The main bottleneck to study this problem arguably is limited data. For training, existing egocentric datasets, such as MultiHop-EgoQA [8] and MM-Ego [42], are derived mostly from the human narrations in Ego4D [15]. While these narrations provide valuable human supervision, they bias toward describing human–object interactions and lack broader environmental context or the diversity of auditory signals in egocentric recordings. For evaluation, existing benchmarks [10, 25, 30] focus mainly on visual cues, limiting their ability to assess integrated audio–visual reasoning. Though exocentric audio–visual benchmarks [21, 24, 35, 41] exist, their multi-modal dynamics are fundamentally different.

To address these problems, we introduce EgoAVU, a fully automated data engine that can generate diverse and high quality audio-visual-language data from public egocentric datasets such as Ego4D [15]. EgoAVU comprises four key components: (i) *Enhancing egocentric narrations* by enriching human descriptions with environmental context, visual object details, and audio captions generated using diverse open-source MLLMs [3, 4, 38]; (ii) *Filtering videos with rich audio–visual dynamics* by selecting egocentric clips featuring varied action sequences, human–object interactions, and a wide range of ambient and foreground sounds; (iii) *Generating fine-grained event captions through audio–visual correlation* by integrating modality-specific cues such as actions, objects, and sounds, and modeling their relationships to enhance multimodal reasoning; and (iv) *Curating diverse audio–visual understanding tasks* that span grounding, temporal reasoning, scene understanding, and audio–visual hallucination.

Leveraging EgoAVU, we construct EgoAVU-*Instruct* — a large scale training dataset of 9K egocentric videos with 3M audio–visual-language samples, and EgoAVU-*Bench* — an evaluation split of 900 videos containing 3K manually verified samples. The generated data features strong multi-modal correspondence, long average video durations (4min), open and close-ended questions capturing diverse aspects of egocentric audio-visual understanding.

Comprehensive experiments show that existing MLLMs perform poorly on EgoAVU-*Bench*, revealing significant limitations in their joint audio–visual reasoning capabilities. The seven models we tested show consistent bias towards the vision modality, often neglecting audio cues or struggle to connect them to the correct visual source. Fine-tuning MLLMs such as Qwen2.5-Omni [38] on EgoAVU-*Instruct* can effectively close this gap, resulting in up to 113% relative performance boost on EgoAVU-*Bench*. More importantly, the performance gain is transferrable to other egocentric benchmarks, such as EgoTempo [30] and EgoIllusion [34], achieving up to 28% relative performance improvement.

2. Related Work

Multimodal-Large Language Models. Recent advances in Multimodal Large Language Models (MLLMs) [2, 11, 29, 36–39] have substantially extended the capabilities of large language models [2, 3, 40] beyond text, enabling unified understanding of visual and auditory inputs. Despite this progress, existing MLLMs face critical limitations when applied to egocentric audio-visual understanding. First, leading models developed for egocentric video understanding, such as MM-Ego [42] and EgoVLPv2 [31], lack the ability to incorporate audio cues, limiting them to perform visual only tasks. Second, even models capable of handling viusal and auditory signals, including Qwen2.5-Omni [38] and Video-LLaVA2 [11], remain primarily trained and benchmarked on exocentric audio-visual data [14, 41]. As a result, they struggle to generalize to egocentric settings, which exhibit fundamentally different characteristics such as dynamic camera motion, frequent self-occlusions, and distinct audio profiles. These limitations highlight a *significant gap in current MLLMs' to jointly understand audio–visual cues in egocentric videos.*

Egocentric-Video Understanding. The field of egocentric video understanding has gained increasing attention due to its relevance in augmented reality (AR) [26, 27] and embodied AI [32, 33]. Large-scale datasets such as Ego4D [15], EPIC-KITCHENS [28], and Ego-Exo4D [16] have driven progress by enabling the creation of egocentric video–language corpora (e.g., MultiHop-EgoQA [8], EgoTextVQA [45], QaEgo4D [6]) and benchmarks (e.g., EgoSchema [25], EgoTempo [30], EgoIllusion [34]). However, these datasets are predominantly constructed from textual narrations, focusing on human–object interactions while overlooking environmental and auditory cues, critical for accurately inferring actions, contextual dynamics, and scene semantics in egocentric settings. Recent benchmarks such as EgoTempo and EgoIllusion attempt to enhance contextual understanding through visual captioning, but their reliance on closed-source models (e.g., Gemini [36], GPT-4o [2]) makes large-scale, reproducible data generation challenging. In this work, we address these problems by *introducing a scalable data-engine capable of generating*

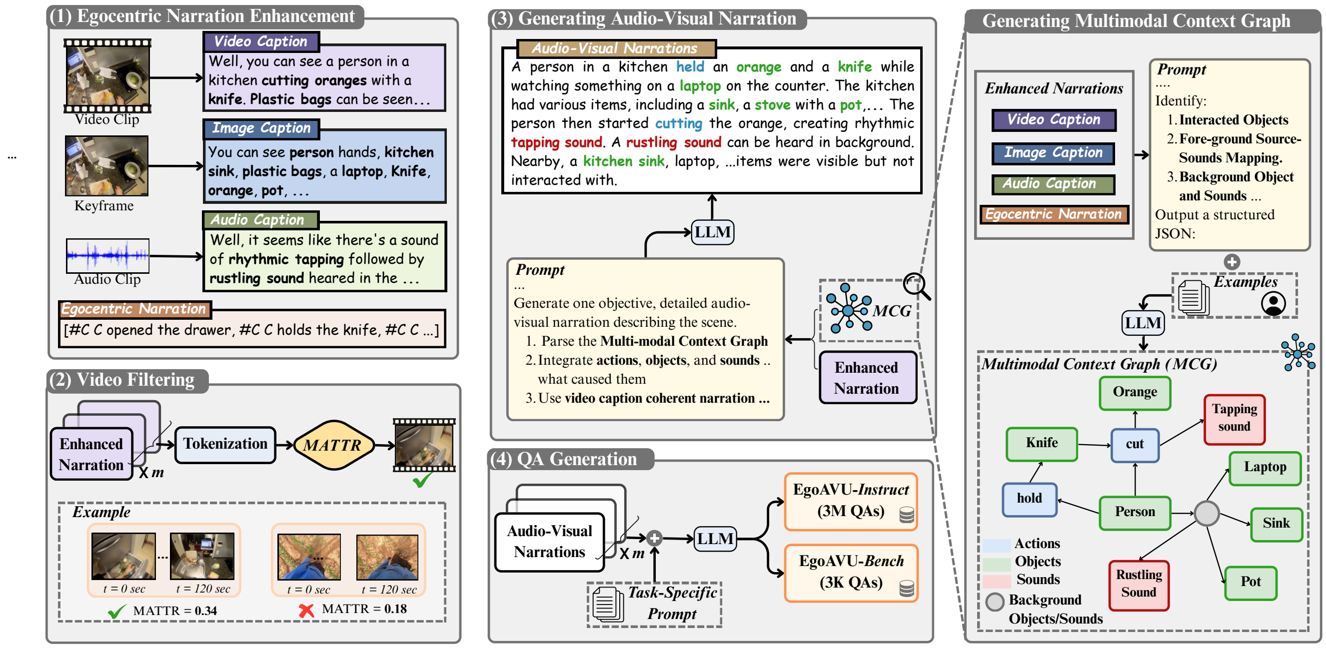

Figure 2. **EgoAVU pipeline.** EgoAVU consists of four key components. (1) For each egocentric video clip, EgoAVU enhances the raw narration with detailed multisensory context using open-source MLLMs [4, 38]. (2) These enriched narrations are then used to select clips that exhibit diverse audio–visual dynamics. (3) Next, EgoAVU constructs a Multimodal Context Graph (MCG), automatically generated via open-source LLMs [3], to capture complex cross-modal relations. The MCG is parsed alongside the enhanced narrations to produce coherent audio–visual narrations. (4) The generated audio-visual narrations are leveraged to generate high-quality audio–visual QA pairs, forming both the instruction-tuning dataset EgoAVU-*Instruct* and the evaluation benchmark EgoAVU-*Bench*.

time-aligned joint audio-visual generation while utilizing open-source MLLMs [3, 4, 38].

3. Method

Fig.2 provides an overview of EgoAVU. We begin by collecting and organizing egocentric data (Sec.3.1). Next, we enhance egocentric narrations using MLLM-generated multi-modal context (Sec.3.2), and use the tokens of enriched narration to filter videos for temporal diversity (Sec.3.3). We then introduce a multi-modal context graph (MCG) that captures complex cross-modal relations, and parse these graphs with open-source LLMs to fuse multi-modal information into a single dense narration (Sec.3.4). Finally, this fused narration is used to generate QA pairs for joint audio-visual understanding (Sec.3.5). To ensure the scalability, EgoAVU only utilizes open source models. Due to the space limit, we provide example outputs and prompts for each module in Appendix A.

3.1. Data Collection

We begin by collecting videos from the Ego4D dataset [15], and filter out videos that lack audio tracks. Each video in Ego4D is accompanied by a set of narrations that provide first-person descriptions of the events, such as "#C C holds a cup," where "#C C" refers to the camera wearer. These narrations can be represented as $\{N_j, t_j\}_{j=1}^{K}$, where N_j denotes the narration text, t_j is the corresponding timestamp, and K is the total number of narrations in the video. To obtain the temporal boundries of each narration, we use the strategy proposed in prior works [23, 30], where temporal boundries T_j for narration N_j is defined as:

$$T_j = \left\{ t_j - \frac{\beta_i}{2\alpha}, t_j + \frac{\beta_i}{2\alpha} \right\}. \quad (1)$$

β_i represents the average interval between consecutive timestamps ($\{T_j, T_{j+1}\}$), and α denotes the global average of β_i across the entire dataset. Since the segments associated with each narration are short (on average 3s), they are insufficient to capture fine-grained visual and auditory information. Following [13], for each video v, we group consecutive segments and their corresponding narrations to form video clips v_j of at least 10s and not more than 360s.

3.2. Narration Enhancement

After data collection, we generate descriptions about audio and video frames to enrich the original egocentric action-centric narrations. The goal of this stage is to obtain detailed descriptions for both modalities, which paves the way for diverse and fine-grained QA generation. The most straightforward approach is to input both video frames and audio tracks into an MLLM, and prompt it to generate the audio and visual descriptions. However, our initial experiments show that MLLMs, when provided with audio and video inputs together, cannot capture important details due to modality bias and hallucination.

Specifically, we evaluate open-source models, including Qwen2.5-Omni [38] and MiniCPM-o [29], on 200 randomly sampled video clips. Each model is first used in a uni-modal setting, captioning visual and auditory information independently, i.e., visual captioning without audio input and audio captioning without video input. We then perform joint audio–visual captioning by providing both modalities simultaneously. After that, we manually compare the consistency for both visual and audio modalities by computing the ratio where an object/event captured in the uni-modal output is also correctly appearing in the multi-modal output. I.e., whether the multi-modal output can capture all the details as in the uni-modal settting.

We observe a consistent pattern where models either omit various sounds or associate audio cues with incorrect visual events. For example, Qwen2.5-Omni shows an error rate of 54.3% for audio and 25.4% for visual consistency, while MiniCPM-o yields 68.2% and 31.2%, respectively (see Appendix C for examples). These findings show that existing MLLMs struggle to maintain accurate and detailed cross-modal grounding in egocentric contexts, motivating a modular data engine that leverages specialized models to process different modalities independently.

With this observation, we leverage a collection of MLLMs to extract rich multisensory information from individual modalities separately. Specifically, for each video segment, we first capture detailed spatial descriptions of objects by applying an image captioner such as Qwen2.5-VL [4] to the center frame. To model temporal dynamics, including camera motion, action sequences, and auditory events, we employ Qwen2.5-Omni in two complementary modes. First, we use it as a video captioner, processing video frames without audio, to generate coherent video-level narrations. Then, we use the same model as an audio captioner without visual inputs to produce detailed auditory narrations that capture both foreground sounds closely tied to human activities (e.g., impacts and hissing) and background sounds (e.g., bird chirping, wind blowing). This process results in time-aligned uni-modal narrations that encapsulate both visual and auditory aspects of the scene.

3.3. Video Filtering for Diversity Enhancement

After obtaining the enhanced narrations, we use them to compute lexical diversity, filtering videos with rich and diverse visual and auditory signals. Specifically, for each video v, we combine the segment-level narrations into a single narration that captures all objects, actions, and sounds present in the video. We then tokenize this narration into $\mathcal{T}_v = \{t_1, \dots, t_n\}$ and compute the Moving-Average Type–Token Ratio (MATTR) [12], which averages the proportion of unique words across sliding windows of size w:

$$\text{MATTR}(\mathcal{T}_v) = \frac{1}{n-w+1} \sum_{i=1}^{n-w+1} \frac{|\text{Uni.}(t_i, \dots, t_{i+w-1})|}{w}. \tag{2}$$

Higher MATTR scores indicate narrations that describe a wider variety of objects, actions, and auditory events. We retain videos whose MATTR exceeds a threshold $\tau = 0.3$, effectively removing the bottom 25% of our distribution to filter static or repetitive descriptions, leading to the final video count of 9,900.

3.4. Audio-Visual Narration Generation

Our next objective is to generate a unified narration that captures both auditory and visual information for each segmented video clip. We do this leveraging text-based LLMs. When synthetically merging unimodal narrations, the LLM must first implicitly retrieve key details across modalities, such as which objects the human interacts with, which remain in the background, and what actions or objects produce specific sounds. It must then integrate this information into a coherent and perceptually grounded narration. However, our preliminary experiments show that directly prompting open-source LLMs such as LLaMA-70B [3] often fails to maintain consistent human–object interactions and sound–source associations in their responses (See Appendix A for detailed comparison).

We design a two-stage pipeline that addresses this challenge. First, we organize audio-visual cues from uni-modal narrations into a structured representation called the *Multi-modal Context Graph (MCG)*. As shown in Fig. 2, MCG captures relationships between visible actions, objects, and audible sounds. Each MCG is generated by prompting an LLM (LLaMA-70B) with enhanced narrations to extract the following information :

- *Interacted Objects:* Objects with which the person physically interacts, along with interaction types, inferred from action narrations.
- *Background Objects:* Objects visible in the environment but never interacted with, identified by comparing objects and action narrations.
- *Foreground Sounds:* Human-induced action sounds correlated with specific actions (e.g., *impact sound* → "C places the phone on the table"), or ambient sounds grounded in visible scene elements (e.g., "dog barking" aligned with a visible dog).
- *Background Sounds:* Sounds present in the audio track whose sources cannot be visually grounded.

We find that MCG makes multi-modal relationships explicit and easily inferable, which we utilize further to generate joint audio-visual narrations.

Category	Example
	Open-Ended
Source–Sound Association	What is the source of the sizzling sound heard in the video?
Audio–Visual Dense Narration	Describe in detail what the person sees, hears, and does throughout the video.
Audio–Visual Segment Narration	Between 240 and 250 seconds, describe the person's surroundings, actions, and the sounds that can be heard?
	Close-Ended
Temporal Reasoning	What happened before the person opened a kitchen cabinet?. Choose the correct option from the following options: ...
Audio–Visual Hallucination	Is there a beeping sound coming from the microwave in the video?

Table 1. **Question Prompts.** Examples of open-ended and close-ended questions in EgoAVU-*Instruct* and EgoAVU-*Bench*.

Dataset	A&V	# Test	Avg. Dur.	Q-Type	# Ans.
EgoTaskQA [19]	✗	8k	25s	OpenQA	13
EgoSchema [25]	✗	500	3 min	CloseQA	-
EgoThink [10]	✗	750	45s	OpenQA	4
EgoTempo [30]	✗	500	45s	OpenQA	10
EgoIllusion [34]	✗	8k	3 min	Open / CloseQA	3
EgoAVU-*Bench*(Ours)	✓	3k	4 min	Open / CloseQA	200

Table 2. **Benchmark statistics.** Beside having QAs with high audio-visual coherence, EgoAVU-*Bench* features large number of QA pairs, longer egocentric videos with audio track, support both open/closed QA type and consists of significantly longer and descriptive responses.

Next, we leverage MCGs to guide the generation of unified, high-coherence audio–visual narrations. The goal is to fuse both visual and auditory details into the same multi-modal coherent narration. As shown in Fig 2, we provide the LLM (LLaMA-70B) with enhanced narrations and MCGs, and prompt it to generate the combined narration. The prompt asks the LLM to first extract explicit cues from the MCG, including interacted and background objects, grounded sound events, and their associations with actions and visible sources, and then align these cues with the corresponding temporal descriptions from the video and action-level narrations.

3.5. QA Generation

Using the unified audio–visual narrations, EgoAVU synthesizes question–answer pairs that probe diverse aspects of egocentric audio-visual understanding.

EgoAVU Taxonomy. We design five categories of QAs encompassing both *open-ended* questions, where models must produce descriptive responses integrating visual and auditory cues, and *closed-ended* questions, which test precise

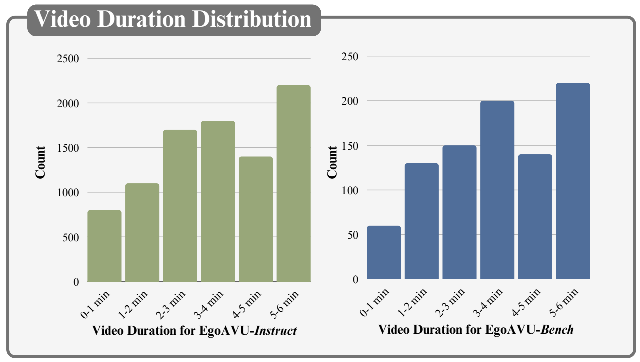

Figure 3. **Video duration distribution.**

multi-modal perceptual understanding through multiple-choice or binary (Yes/No) formats. A detailed description of each category is provided below.

- **Open-ended QAs.** These questions are primarily free-form, requiring the model to develop a comprehensive understanding of temporal and spatial dynamics across visual and auditory cues in egocentric videos. The specific categories include: (1) *Sound–Source Association (SSA):* Identify various foreground sounds in the video and determine their corresponding visible sources including human actions or various objects shown in the video. (2) *Audio–Visual Segment Narration (AVSN):* Answer segment-level questions by producing coherent and natural narrations that describe what the person is doing, seeing, and hearing, including both foreground and background sounds, grounded within the specified temporal range. (3) *Audio–Visual Dense Narration (AVDN):* Extend previous task to the entire video, assessing the model's ability to maintain narrative coherence across the complete video.
- **Close-ended QAs.** These questions are particularly useful for *adversarial testing*. Close-ended QAs allow the construction of fine-grained distractors and counterfactuals to probe a model's susceptibility to spurious correlations in audio-visual understanding. We design two main categories: (1) *Temporal Reasoning (TR):* Multiple-choice questions with four options that assess the model's understanding of temporal relationships among multi-modal events such as human actions, visual objects, and auditory cues in egocentric videos. These questions focus on reasoning about event ordering, for example identifying which event occurred first or last, or answering before/after queries. (2) *Audio–Visual Hallucination (AVH):* Binary ("Yes"/"No") questions that evaluate the model's tendency to hallucinate when detecting visual or auditory signals, such as verifying the presence of actions, objects, or sounds throughout the video.

Models	Size	SSA	AVDN			AVSN			TR	AVH
		S (↑)	**S** (↑)	**M** (↑)	**R** (↑)	**S** (↑)	**M** (↑)	**R** (↑)	**Acc.** (↑)	**Acc.** (↑)
Open-Source MLLMs										
VideoLLaMA2 [11]	7B	1.51	1.88	3.65	8.32	1.71	7.50	13.89	37.00	20.32
Baichuan-Omni [20]	7B	1.49	1.92	4.82	9.75	1.79	8.34	14.21	39.85	21.10
Intern-Omni [9]	8.7B	1.47	1.95	5.20	10.11	1.82	8.69	14.70	41.22	21.75
Phi4-mm [1]	8B	1.42	1.59	8.79	13.13	1.69	12.17	16.90	45.04	22.89
MiniCPM-o [29]	8B	1.43	2.27	10.84	14.77	2.06	9.68	12.19	26.44	21.76
Qwen2.5-Omni [38]	3B	1.45	2.17	8.55	13.45	1.85	8.63	13.08	46.40	26.28
Qwen2.5-Omni [38]	7B	1.50	2.37	10.69	14.74	1.99	9.99	13.39	53.20	42.69
MLLM trained with EgoAVU-Instruct										
OURS $_{(LoRA)}$	7B	3.15	2.60	12.20	17.19	2.45	22.53	28.34	64.31	**61.69**
OURS $_{(Full)}$	7B	**3.20**	**2.66**	**12.50**	**17.32**	**2.63**	**22.68**	**28.70**	**67.84**	60.12
Δ(%)	–	+113.3	+12.2	+16.9	+17.2	+27.6	+86.5	+69.8	+27.2	+30.8

Table 3. **Main result on EgoAVU-*Bench*.** We compare seven MLLMs with joint audio–visual understanding capabilities against our fine-tuned models across a diverse set of tasks in EgoAVU-*Bench*, including open-ended QAs: Source-Sound Association (SSA), Audio-Visual Dense Narration (AVDN), and Audio-Visual Segment Narration (AVSN), as well as closed-ended QAs: Temporal Reasoning (TR) and Audio-Visual Hallucination (AVH). For the open-ended tasks, we report LLM-as-Judge (S), METEOR (M), and ROUGE-L (R). For the closed-ended tasks, we report Accuracy (Acc.). Additionally for each task, we compute the relative performance gain (Δ) between the best open-source model and our fine-tuned models.

3.6. Dataset Overview

Leveraging EgoAVU, we construct the first egocentric audio–visual dataset suite for training and evaluating MLLMs. Our large-scale instruction-tuning corpus, EgoAVU-*Instruct*, comprises approximately 3M samples with 9K egocentric videos, while the evaluation benchmark, EgoAVU-*Bench*, contains 3K QA pairs with 900 distinct videos. Both EgoAVU-*Instruct* and EgoAVU-*Bench* encompass a wide range of real-world visual scenarios, such as cooking, painting, and other indoor/outdoor activities (see Appendix D for detailed distributions). Extensive manual verifications are conducted on EgoAVU-*Bench* to ensure that the audio-visual information is correctly grounded. As illustrated in Fig. 3, both EgoAVU-*Instruct*and EgoAVU-*Bench* covers all five question categories introduced earlier, featuring a balanced mix of open- and closed-ended formats. As shown in Fig. 3, the video length in both datasets vary from 1 to 6 minutes, providing rich temporal diversity for model training and evaluation.

Table 2 compares existing egocentric benchmarks with our EgoAVU-*Bench*. Besides filling the gap of egocentric audio-visual understanding, EgoAVU-*Bench* includes a large number of QAs, features longer videos with synchronized audio tracks, supports both open- and closed-ended question formats, and provides significantly longer and more descriptive responses.

4. Results

Implementation Details. We verify the effectiveness of our instruction-tuning dataset, EgoAVU-*Instruct*, by fine-tuning MLLMs such as Qwen2.5-Omni (7B) on it and compare the finetuned model with baselines on EgoAVU-*Bench*. Fine-tuning is performed using LLaMA-Factory [44], under two settings: LoRA [18] and full fine-tuning. All experiments are conducted on 64 H100 GPUs with a global batch size of 64, training each model for 5 epochs. To ensure consistent visual coverage across samples, during training, we uniformly sample 300 frames per video. We also uniformly sample data from each of the 5 tasks to achieve a balanced performance (See Appendix E for additional details).

Evaluation Protocol. For close-ended QAs in EgoAVU-*Bench*, we follow [43] and use regex-based string matching, where we construct robust regular expressions and design a response-processing module to extract key phrases such as option IDs (A, B, C, D), binary indicators (yes/no), and conclusion phrases from long responses for accurate answer matching. For open-ended QAs, similar to prior work [30], we adopt the LLM-as-a-judge approach, employing Qwen3-235B-A22B-Instruct-2507 [40] as an open-source judge for reproducibility. The model rates MLLM-generated responses on a 1–5 scale (see Appendix E for the evaluation prompt). We additionally report standard metrics used for dense response evaluation, including ROUGE-L [22] and METEOR [5].

Model	EgoTempo [30] Acc. (↑)	EgoSchema [25] Acc. (↑)	EgoIllusion [34] Acc. (↑)
Qwen2.5-Omni	16.25	**67.43**	56.32
Ours $_{(LoRA)}$	**20.83**$_{+28.1\%}$	67.34$_{-0.1\%}$	**60.36**$_{+7.2\%}$
Ours $_{(Full)}$	20.21$_{+24.4\%}$	66.21$_{-1.8\%}$	60.24$_{+7.0\%}$

Table 4. **Result on other Egocentric benchmarks.** Finetuning on EgoAVU-*Instruct* also benefits other egocentric benchmarks.

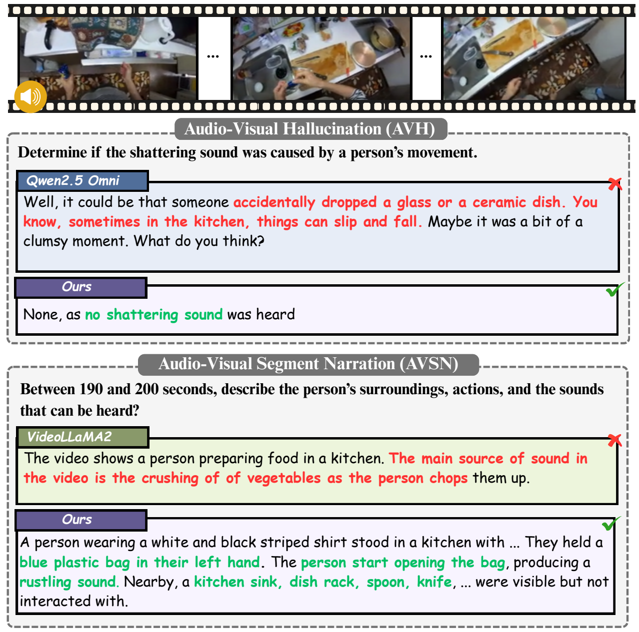

Figure 4. **Qualitative Analysis on EgoAVU-*Bench*.** We observe existing MLLMs, such as Qwen2.5-Omni and VideoLLaMA2, exhibit *strong visual bias* when answering audio-visual questions from EgoAVU-*Bench* (highlighted in red). In contrast, our model demonstrates superior *sound–source coherence* and a better understanding of *action sequences* (highlighted in green)

4.1. Main Results

Table 3 presents the main result on EgoAVU-*Bench*. The key findings are summarized below.

MLLMs struggle to associate sounds with their visual sources. This is evident from their low scores on the Source-Sound Localization (SSA) task, where LLM-as-a-judge (S) evaluations remain below 1.6 out of 5 across all baseline MLLMs.

MLLMs struggle to produce coherent and temporally aligned audio-visual narrations. In both the Audio-Visual Dense Narration (AVDN) and Audio-Visual Segment Narration (AVSN) tasks, even the best-performing model, Qwen2.5-Omni (7B), attains S scores of below 2.4, which is consistent with low caption quality metrics such as ROUGE-L (R) and METEOR (M).

MLLMs demonstrate limited temporal reasoning over joint audio-visual inputs. The highest Temporal Reasoning (TR) accuracy, achieved by Qwen2.5-Omni (7B), is merely 53.2%, with models such as VideoLLaMA2 and MiniCPM-o performing considerably worse.

MLLMs frequently hallucinate during audio-visual reasoning, as reflected by their below 43% accuracy on the Audio-Visual Hallucination (AVH) task.

Finetuning MLLMs on EgoAVU-Instruct effectively improves egocentric audio-visual understanding. Our finetuned model yields substantial and consistent performance gains across all tasks. Compared to the best performing baseline, we achieve up to 113.3% and 44.5% relative performance improvement on open and close ended tasks respectively. In addition, LoRA and full finetuning both provides considerable performance gain. This shows the possibility to achieve strong audio-visual understanding under resource-limited scenarios.

Figure 4 compares the responses of VideoLLaMA2 and Qwen2.5-Omni (7B), with our fine-tuned model on both open-ended and close-ended tasks. In close-ended tasks like Audio-Visual Hallucination, when queried about a sound source absent from the video, Qwen2.5-Omni often fabricates a visually plausible yet non-existent source, whereas our model effectively resists such misleading prompts through improved audio–visual grounding. For open-ended tasks such as Audio-Visual Segment Narration, our model exhibits stronger sound–source coherence and more accurate action-sequence understanding. (See Appendix E for additional examples.)

4.2. Evaluating Existing Egocentric Benchmarks

To further evaluate the generalizability of our fine-tuned model, we report its performance on additional egocentric video–language benchmarks, including EgoTempo, EgoSchema, and EgoIllusion. As shown in Table 4, finetuning on EgoAVU-*Instruct* leads to notable accuracy improvements on EgoTempo and EgoIllusion, with gains of up to 28.1%, while maintaining competitive performance on EgoSchema, showing only a marginal decrease of 0.1%. Note that LoRA performs slightly better than full finetuning in terms of improving other datasets. These results demonstrate that training on EgoAVU-*Instruct* enhances audio-visual understanding without causing overfitting, and further complements performance across diverse egocentric video QA tasks. We also evaluate our finetuned model on non-egocentric datasets. As shown in Appendix F, finetuning MLLMs on EgoAVU-*Instruct* can maintain the performance on non-egocentric datasets.

4.3. Error Analysis

Intrigued by the weak performance of current MLLMs across all tasks in EgoAVU-*Bench*, we conduct a detailed error analysis separately for the close-ended and open-ended tasks to analyze their behavior patterns. Our key find-

Models	Size	Action Acc. (↑)	Object Acc. (↑)	Sound Acc. (↑)
VideoLLaMA2	7B	21.36	20.32	19.27
Phi4-MM	8B	22.45	21.74	24.49
MiniCPM-o	8B	24.49	24.46	16.33
Qwen2.5-Omni	3B	25.00	29.35	24.49
Qwen2.5-Omni	7B	44.39	50.00	33.67
OURS $_{(LoRA)}$	7B	60.20	61.09	63.78
OURS $_{(Full)}$	7B	61.32	62.40	64.20

Table 5. **Error Analysis for Audio-Visual Hallucination (AVH).**

Models	Size	Action Acc. (↑)	Object Acc. (↑)	Sound Acc. (↑)
VideoLLaMA2	7B	45.88	38.38	38.30
Phi4-MM	8B	48.81	48.48	46.74
MiniCPM-o	8B	40.48	25.00	37.50
Qwen2.5-Omni	3B	44.44	69.70	31.91
Qwen2.5-Omni	7B	43.53	64.65	36.17
OURS $_{(LoRA)}$	7B	54.31	68.90	52.45
OURS $_{(Full)}$	7B	55.29	69.80	53.17

Table 6. **Error Analysis for Temporal Reasoning (TR)**

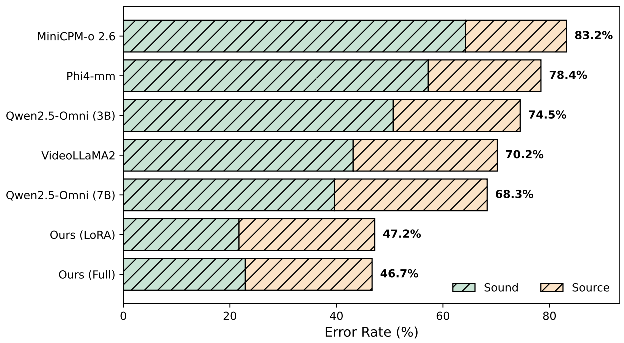

Figure 5. **Error Analysis on Sound-Source Association (SSA).** For open-ended tasks, MLLMs with high error rates primarily fail to perceive sounds rather than identifying visual objects/events. For less advanced models, identifying sound is a heavier issue.

ings are summarized below:

Close-Ended Tasks. Firstly, as illustrated in Table 5 and 6, for close-ended tasks such as Audio-Visual Hallucination (AVH) and Temporal Reasoning (TR), we evaluate MLLMs ability to independently perceive multisensory inputs within egocentric videos, specifically, human actions, visual objects, and sounds, including both foreground and background audio cues. This is achieved by separately evaluate the accuracy on the 3 subsets of corresponding tasks. Overall, our analysis reveals a consistent trend: *MLLMs struggle the most with identifying sounds, followed by human actions, while performing relatively better at recognizing visual objects.* For instance, in the TR task, the best-performing model, Qwen2.5-Omni, achieves only 36.1% accuracy in identifying sounds, exhibiting substantial performance gaps of 28.5% and 7.4% relative to visual object and human action identification, respectively. Furthermore, we find that model fine-tuned on our dataset, achieves a significant boost in its ability to independently perceive multisensory inputs within egocentric videos. For example, in the AVH task, compared to Qwen2.5-Omni, our fine-tuned model shows a reduced hallucination rate of 15.9%, 11.0%, and 30.0% when identifying human actions, visual objects, and sounds, respectively.

Open-Ended Tasks. We further extend our analysis to open-ended tasks such as Source-Sound Association (SSA) in EgoAVUBench, aiming to determine which modality contributes more to the overall error rate when models are asked to produce joint audio–visual descriptions. Specifically, we randomly sample 200 data points from the SSA task and, for each incorrect response across different MLLMs, manually annotate whether the error stems from inaccurate sound perception or an incorrect source description (which may involve visible objects or human–object interactions in the egocentric video). As shown in Fig. 5, our model achieves a substantially lower error rate of 21.1% compared to the next best model, Qwen2.5-Omni. Interestingly, other MLLMs such as MiniCPM-o and Phi4-mm, which exhibit much higher error rates, show that over 72% of their errors stems from incorrect or missed sound descriptions rather than misidentified human–object interactions. This indicates that *MLLMs primarily fail due to their limited ability to accurately perceive and interpret sounds, leading to incorrect associations with their visible sources.*

5. Conclusion

We presented EgoAVU, a novel data engine to enhance egocentric audio-visual understanding by addressing the data limitations. EgoAVU employed modularized MLLMs to enhance egocentric narrations, and leveraged multi-modal context graphs to generate diverse, high-quality audio-visual QA pairs. Our evaluation benchmark EgoAVU-*Bench* revealed for the first time that existing MLLMs exhibit consistent vision bias, often neglecting or hallucinating audio information in egocentric videos. Our large-scale training dataset EgoAVU-*Instruct* effectively mitigated this gap, significantly improved performance on both EgoAVU-*Bench* and existing egocentric benchmarks. Importantly, EgoAVU demonstrated the self-learning potential of MLLMs: using uni-modal capabilities to improve the joint-modal capability. In terms of limitation, our training data, though with carefully designed filtering techniques, still contains noise from open source MLLM outputs. We believe this problem can be continually alleviated as the uni-modal capabilities of MLLMs improve, and leave the development of more soficsticated approaches as future work.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 6

[2] 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. 2

[3] AI@Meta. Llama 3 model card. 2024. 2, 3, 4

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 2, 3, 4

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[6] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 2

[7] Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos, 2024. 2

[8] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2159–2167, 2025. 2

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6

[10] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024. 2, 5

[11] 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, 6

[12] Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010. 4

[13] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2024. 3

[14] Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18959–18969, 2025. 2

[15] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 2, 3

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2

[17] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Imagebind-llm: Multi-modality instruction tuning, 2023. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[19] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 5

[20] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024. 6

[21] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024. 2

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 6

[23] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 3

[24] Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. *Advances in Neural Information Processing Systems*, 37:9507–9531, 2024. 2

[25] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2, 5, 7

[26] Nuno Cid Martins, Bernardo Marques, Paulo Dias, and Beatriz Sousa Santos. Extending the egocentric viewpoint in situated visualization using augmented reality. In *2023 27th International Conference Information Visualisation (IV)*, pages 83–89. IEEE, 2023. 2

[27] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. *Advances in Neural Information Processing Systems*, 36: 60130–60143, 2023. 2

[28] Amirshayan Nasirimajd, Simone Alberto Peirone, Chiara Plizzari, and Barbara Caputo. Epic-kitchens-100 unsupervised domain adaptation challenge: Mixed sequences prediction. *arXiv preprint arXiv:2307.12837*, 2023. 2

[29] OpenBMB. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2024. Accessed: 2025-03-07. 2, 4, 6

[30] Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24129–24138, 2025. 2, 3, 5, 6, 7, 12

[31] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 2

[32] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020. 2

[33] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 2

[34] Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. EGOILLUSION: Benchmarking hallucinations in egocentric video understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28449–28468, Suzhou, China, 2025. Association for Computational Linguistics. 2, 5, 7

[35] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024. 2

[36] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[38] 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, 4, 6

[39] 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

[40] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 6

[41] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. 2

[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. *arXiv preprint arXiv:2410.07177*, 2024. 2

[43] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 6

[44] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 6

[45] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373, 2025. 2