

# When Eyes and Ears Disagree: Can MLLMs Discern Audio-Visual Confusion?

Qilang Ye<sup>1 2</sup>, Wei Zeng<sup>2</sup>, Meng Liu<sup>2 3\*</sup>, Jie Zhang<sup>4</sup>, Yupeng Hu<sup>6</sup>, Zitong Yu<sup>4 5\*</sup>, Yu Zhou<sup>1 2</sup>

<sup>1</sup>VCIP & TMCC & DISSec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University

<sup>2</sup>Zhongguancun Academy

<sup>3</sup>School of Computer Science and Technology, Shandong Jianzhu University

<sup>4</sup>School of Information Science and Technology, Great Bay University

<sup>5</sup>Dongguan Key Laboratory for Intelligence and Information Technology

<sup>6</sup>School of Software Engineering, Shandong University

{s-yql25, s-zw24}@bjzgca.edu.cn, mengliu.sdu@gmail.com, jz@stu.cqut.edu.cn, huyupeng@sdu.edu.cn, yuzitong@gbu.edu.cn, yzhou@nankai.edu.cn

## Abstract

*Can Multimodal Large Language Models (MLLMs) discern confused objects that are visually present but audio-absent?* To study this, we introduce a new benchmark, AV-ConfuseBench, which simulates an “Audio-Visual Confusion” scene by modifying the corresponding sound of an object in the video, e.g., mute the sounding object and ask MLLMs “Is there a/an {muted-object} sound?”. Experimental results reveal that MLLMs, such as Qwen2.5-Omni and Gemini 2.5, struggle to discriminate non-existent audio due to visually dominated reasoning. Motivated by this observation, we introduce **RL-CoMM**, a **R**einforcement **L**earning-based **C**ollaborative **M**ulti-**M**LLM that is built upon the Qwen2.5-Omni foundation. RL-CoMM includes two stages: 1) To alleviate visually dominated ambiguities, we introduce an external model, a Large Audio Language Model (LALM), as the reference model to generate audio-only reasoning. Then, we design a Step-wise Reasoning Reward function that enables MLLMs to self-improve audio-visual reasoning with the audio-only reference. 2) To ensure an accurate answer prediction, we introduce Answer-centered Confidence Optimization to reduce the uncertainty of potential heterogeneous reasoning differences. Extensive experiments on audio-visual question answering and audio-visual hallucination show that RL-CoMM improves the accuracy by 10~30% over the baseline model with limited training data.

**Code** — <https://github.com/rikeilong/AVConfusion>

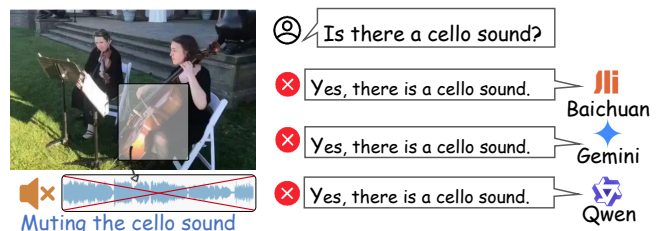
## Introduction

We perceive things by gathering visual information from our eyes while constantly acquiring knowledge through our hearing. Recent advancements in Multimodal Large Language Models (MLLMs<sup>1</sup>) (Ye et al. 2024; Cheng et al. 2024; Zhang, Li, and Bing 2023) show a remarkable ability to understand real-world human language and generate continuum contexts. Furthermore, the wide range of complex

\*Corresponding Authors

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

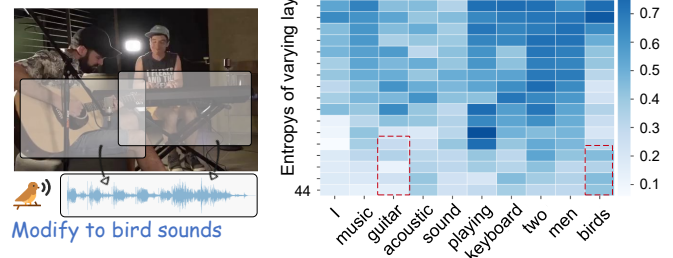
<sup>1</sup>In this paper, MLLMs refer to models that can process inputs containing both video and audio.



(a) Error cases in audio-muted scenarios.

**Query:** Describe what you hear?

**Qwen2.5-Omni:** I hear two men playing music together. **One is playing an acoustic guitar, and the other is playing a keyboard.** There is also a fireplace in the background, and the music is accompanied by the sound of birds chirping.



(b) Uncertainty of different layers in audio-modified scenarios.

Figure 1: Examples of MLLMs confronting audio-visual confusion.

visual and audio signals has facilitated the development of MLLMs such as Qwen2.5-Omni (Xu et al. 2025) and Gemini 2.5 Pro (Reid et al. 2024). After fine-tuning with an extensive synchronized video and sound corpus, MLLMs develop a cognitive capacity for audio-visual understanding in areas including automatic speech recognition (OpenAI 2023), audio-visual captioning (Zhan et al. 2024), and general audio-visual processing (Han et al. 2023). Despite the excellent generative capabilities of MLLMs for generic scenarios, studies on Audio-Visual Hallucinations (AVH) (Gong et al. 2024) have revealed that MLLMs are unable

to distinguish between the volume and pitch of two pieces of audio. Furthermore, MLLMs are susceptible to hearing fictitious sounds or perceiving imaginary visual objects in cross-modal understanding scenarios (Sung-Bin et al. 2025).

In this paper, we further explore: if the given audio-visual information is asymmetric,  $\diamond$  *Can MLLMs discern confused objects that are visually present but audio-absent?* We call this phenomenon “**Audio-Visual Confusion**”, which refers to the given audio-visual information being inherently asymmetric. As illustrated in Fig. 1, we primarily focus on two settings: (a) **Test whether MLLMs can recognize that a certain muted object is not making a sound.** In normal cognition, vision and audio are bound, i.e., the presence of some object necessitates the presence of sound feedback. But what if given a damaged video that is visually intact but acoustically crippled, would MLLMs recognize it? For example, we input a “cello-muted” video and ask Baichuan-Omni-7B (Li et al. 2024), Gemini 2.5 Pro (Reid et al. 2024), and Qwen2.5-Omni-7B (Xu et al. 2025) if there is a cello sound. All models fail to notice the missing audio but blindly believe in the visual information. Further, we explore (b) **Test whether MLLMs can balance audio-visual information for reasoning.** For example, we input an extremely unbalanced video (tampering with the music as bird sounds) to Qwen2.5-Omni. Then, we follow an entropy-based metric (Farquhar et al. 2024; Zou et al. 2024) to quantify the uncertainty of the next token in each layer. As shown on the right of Fig. 1, although Qwen2.5-Omni is able to distinguish *bird* sounds, its uncertainty scores remain relatively high in the latter layers, which means the model is still being biased towards hearing *acoustic guitar* and *keyboard* sounds when generating responses.

To further investigate this phenomenon, we propose a mini-benchmark, **AV-ConfuseBench**, to assess the ability of MLLMs to distinguish audio-visual confusions. It consists of two settings: audio-muted (i.e., muting one of the multiple instruments all the time) and audio-modified (i.e., tampering with the soundtrack of the entire video to be completely out of sync with the theme). Audio-muted setting is tested by giving a query: “Is there a/an {muted-object} sound?” and models reply yes/no; audio-modified setting aims at evaluating the consistency of the generated content with the audio-visual ground truth.

As shown in Table 1, we find that: 1) None of the MLLMs, especially open-source models, yield the expected results in the audio-muted scenario, which means they struggle to discriminate non-existent audio tokens due to visual influences. 2) While Gemini 2.5 Pro with thinking mode achieves higher performance over the baseline, it remains difficult to think outside of the inherent internal knowledge in multimodal tasks containing video and audio (About 38.36% of the responses remain affected by visually guided thinking and output errors). 3) MLLMs appear to be insensitive to audio content, and the reasoning is dominated by the presented visual in the audio-modified scenario. This means the generated contexts mostly elaborate visual information.

Motivated by this observation, we introduce **RL-CoMM**, a **Reinforcement Learning-based Collaborative Multi-MLLMs** system that is built upon the Qwen2.5-Omni-3B

Model	Audio-muted		Audio-modified	
	Acc. ( $\uparrow$ )	Yes (%)	A-Acc. ( $\uparrow$ )	V-Acc. ( $\uparrow$ )
<i>Open-source omni-models</i>				
Video-LLaMA2-7B	2.73	97.27	0.88	3.88
Baichuan-Omni-7B	5.47	94.53	1.12	4.07
Qwen2.5-Omni-7B	9.59	90.41	1.02	4.32
<i>Close-source omni-models</i>				
Gemini 2.5 Flash	<u>28.76</u>	<u>71.24</u>	2.24	<b>4.78</b>
Gemini 2.5 Pro	<b>68.50</b>	<b>31.50</b>	<b>2.83</b>	<u>4.66</u>
Random Choice	50.00	50.00	-	-

Table 1: Results of different MLLMs on AV-ConfuseBench. We evaluate various open/closed-source in two settings, where “Yes” is the proportion of answer yes among total responses, “A-Acc.”, “V-Acc.” refer to scoring audio and visual response accuracy on a 0-5 scale using GPT-4.

foundation. RL-CoMM consists of two different stages: 1) Step-wise Reasoning Reward function (Step-RR) and 2) Answer-centered Confidence Optimization (Ans-CO). For the first stage, RL-CoMM introduces an external model, i.e., Large Audio Language Models (LALMs), as the reference model to supplement audio evidence. Then, we design Step-RR based on Group Relative Policy Optimization (GRPO), which uses the extra audio knowledge to incentivize the policy model for audio context reasoning and audio-visual correlation reasoning. For the second stage, Ans-CO can reduce the policy model’s uncertainty over its answer predictions to resolve uncertainties arising from potentially heterogeneous reasoning differences. Through extensive experiments on Audio-visual Question Answering (AVQA) tasks and AVH benchmarks, our proposed RL-CoMM improves accuracy by 10~30% over the base-LLM only with around 20% of total training samples.

## Related Work

**Large Audio-Visual Language Models.** Drawing inspiration from the remarkable ability of Large Language Models (LLMs) (Yang et al. 2024a,b; Touvron et al. 2023) to generate coherent language, studies have extended LLMs to other multimodal tasks, e.g., Audio-visual Question Answering (AVQA) (Yun et al. 2021), and human-centric understanding tasks (Nan et al. 2024; Ye and Yu 2024). Primarily, visual LLMs (Ye et al. 2025; Xie et al. 2024; Liu et al. 2018a,b; Lin et al. 2025) emphasize the design of elegant bridging methods. In this work, we focus on Omni-LLMs (Ye et al. 2024; Xu et al. 2025; Cheng et al. 2024), it refers to models that understand both video and audio inputs, which possess multisensory properties that mimic human perception. However, we find that these models trained from synchronized audio-visual data are highly susceptible to complementary modalities.

**Audio-visual Defects in MLLMs.** Hallucination (Li et al. 2023; ?; Gong et al. 2024; Gunjal, Yin, and Bas 2024; Liu et al. 2024; Shu et al. 2025) refers to the generation of imagi-

native textual responses by the model that do not correspond to the input signal. Studies (Sung-Bin et al. 2025) have revealed that such phenomenon is caused by favoring the internal knowledge of the LLM and disregarding the input signal. In this paper, we present another shortage of MLLMs that is similar to hallucinations: “Audio-Visual Confusion”. Specifically, we test the ability of MLLMs to cope with asymmetric audio-visual information.

**MLLMs with Reinforcement Learning.** Reinforcement Learning (RL) from human feedback (Ouyang et al. 2022; MacGlashan et al. 2017), as an early optimization of language models requires significant human annotation and computational resources. Then, on-policy optimization methods such as DPO (Rafailov et al. 2023), and PPO (Schulman et al. 2017), which reward fine-tuned models via computing advantages have achieved excellent outcomes. GRPO (Shao et al. 2024), as the core optimization algorithm of Deepseek-R1, has advanced the intermediate thinking trajectories in response via KL-penalty and reward model. Recent studies (Liu et al. 2025) have focused on unifying Supervised Fine-Tuning (SFT) (Zhang et al. 2023) and RL to effectively improve the sensitivity of LLMs to the final output. This paradigm utilizes high quality offline data and online-optimization to interleave the training of models. Unlike prior methods, we improve GRPO by including a heterogeneous reference model, audio-LLMs (Chu et al. 2024), to complement Omni-LLMs (Xu et al. 2025) thinking knowledge. Such a paradigm significantly improves the performance of base-LLMs on AVQA with limited data.

## AV-ConfuseBench

We believe that analyzing the reliability of existing MLLMs under audio-visual confusion can develop more robust models. We provide detailed descriptions and construction of two different settings in AV-ConfuseBench below:

**Audio-muted Confusion.** This task is set on masking a particular sound source in a scene where multiple instruments are performing, and it assesses whether visual objects affect the audio understanding of MLLMs. All questions are present in the form of: “*This is a video of audio corruption where some instrument sound is muted. question: Is there a/an {muted-object} sound?*” and the ground truth is “No”. We mute the sound sources of the collected **39** videos and yield a total of **73** Q&A pairs. The evaluation metrics are accuracy and model response “Yes” coverage.

**Audio-modified Confusion.** This task is designed to modify background sounds to assess whether audio-generated false information affects the discriminative ability of MLLMs. All questions are present in the form of “*Describe what you see and what you hear*”. We collected **5** different environmental sounds, including sounds of wind, bird, rain, electric drill, and thunder sounds, to tamper with the background sound of **20** videos and yield a total of **100** Q&A pairs. To ensure the quality of the assessment, all ground truths are manually labeled. The evaluation metrics are AI-assisted (DeepSeek-AI et al. 2025) assessment of the accuracy of generated visual and audio contents to ground truths.

---

Algorithm 1: The process of RL-CoMM.

---

**Require:** Policy model  $\pi_\theta$  after warm-up; Reference model  $\pi_{ref}$   
**Ensure:** Optimized policy model  $\pi_\theta$   
**for** each  $i \in [1, N]$  **do**  
    **if** Stage == Step-RR  
        Generate audio reference reasoning paths  $o_{ref} \sim \pi_{ref}$   
        Generate  $G$  audio-visual reasoning paths  $o^i \sim \pi_\theta$   
        Compute rewards  $r_{format}^i, r_{arr}^i, r_{avc}^i$  via Eqs. 2-4  
        Optimize policy model  $\pi_\theta$  with group advantages  $A^i$   
    **else if** Stage == Ans-CO  
        Clip the answer token  $o_{<t}^i$  in  $o^i$   
        Optimize policy model  $\pi_\theta$  with Ans-CO via Eqs. 6  
**end for**  
**return** Policy model  $\pi_\theta$

---

## Methodology

### Preliminary: RL with Verifiable Rewards

The verifiable reward function, e.g., GRPO (Shao et al. 2024), is a direct optimization strategy designed to enhance the reasoning capability of the policy model  $\pi_\theta(\cdot|\cdot)$ . It simplifies the post-training cost by removing the value model, while generating multiple responses  $\{o\}$  for the input prompt  $q$  to measure correctness. Specifically, the GRPO objective is defined as:

$$\mathcal{L}_{GRPO} = \mathbb{E}_{[(q,o) \sim D]} [R(\theta) - \beta \text{KL}[\pi_\theta(o|q) \parallel \pi_{ref}(o|q)]], \quad (1)$$

where  $\pi_{ref}$  is the reference model,  $R(\theta)$  denotes the reward function:  $\min[\frac{\pi_\theta(o|q)}{\pi_{ref}(o|q)} A^i, \text{clip}(\frac{\pi_\theta(o|q)}{\pi_{ref}(o|q)}; 1-1-\epsilon, 1+\epsilon) A^i]$ ,  $A^i$  denotes the advantage of the  $i$ -th response, and the policy model is optimized via updating the parameter  $\theta$  with gradient ascent. Notably, we have removed the KL loss during the training period.

### RL-CoMM

The framework and process of RL-CoMM are shown in Fig. 2 and Algorithm 1, respectively. We treat the LALM as the reference model  $\pi_{ref}$  and the Omni-LLM as the policy model  $\pi_\theta$ , respectively. RL-CoMM consists of two optimization stages. After warm-up, Step-RR is used to optimize reasoning, Ans-CO is used to optimize answers. The multimodal input data consists of video, audio, and two prompts in the specified format.

**Warm-up for Policy Model.** To ensure a stable reasoning of the foundation model, we introduce supervised fine-tuning of a small number of datasets before online RL optimization. During the warm-up phase, the policy model is trained on a given multimodal dataset, which consists of questions, videos, audios, and long responses constructed with audio-visual context. This phase primarily specifies that the output content of the policy model can include step-by-step visual reasoning and audio reasoning.

**Data Preparation.** Following the recent GRPO-style training paradigm (Zhong et al. 2025), we define two multi-

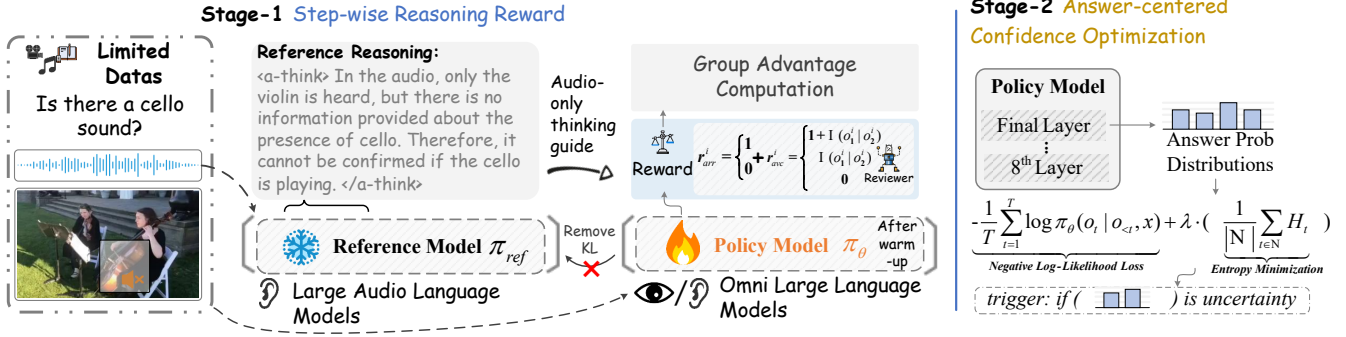


Figure 2: Framework of RL-CoMM, where LALMs serve as the reference model and Omni-LLMs serve as the policy model. Given audio-visual inputs, we first let the LALM generate the reference reasoning for the audio. The policy model is verified by the reviewer (Qwen3 Embedding) to compute group advantages via the Step-wise Reasoning Reward function. Notably, we remove the KL penalty during the policy gradient optimization due to heterogeneous model structure differences. Furthermore, we introduce an Answer-centered Confidence Optimization to reduce uncertainty in the predicted answer of the policy model.

modal training data formats. We define three tags  $\langle a\text{-think} \rangle \langle v\text{-think} \rangle \langle \text{answer} \rangle$ , which serve to sample the audio thinking content, visual thinking content, and predicted answers of the policy models. Notably, we let the reference model generate question-oriented thinking content based on the ground truth. In this way, we avoid the audio-visual correlation thinking inherent to Omni-LLMs and instead reason the question from a specific perspectives. To restrict the output format of the policy model, we use the format reward during the training process as follows:

$$r_{format}^i = \begin{cases} 1, & \text{if format is correct;} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

**Step-wise Reasoning Reward Function.** A crucial step in RL is to design effective reward models, also known as reward functions. It updates the policy model at each step through generalized advantage while aligning the preference algorithms to prevent model changes from differing too much from the reference model. Traditional RL reward designs for Omni-LLMs primarily focus on answer accuracy and format consistency (Xing et al. 2025), lacking incentives for multimodal reasoning content.

For the multiple-choice AVQA tasks, the model should first deduce the audio-visual content, then choose the correct option. We design Step-RR mainly for correcting the potential visual bias and blending audio-visual perception in the reasoning content of Omni-LLMs. Step-RR includes two types of rule-based rewards, i.e., Audio Reasoning Rationality reward (ARR)  $r_{arr}$  and Audio-Visual Correlation (AVC) reward  $r_{avc}$ . Furthermore, to ensure efficient and accurate reward allocation, we use an off-the-shelf lightweight and powerful text embedding model, i.e., Qwen3 Embedding-0.6B (Zhang et al. 2025), for semantic alignment. Specifically, we define the audio reasoning path to the question  $q$  from the reference model  $\pi_{ref}$  as  $o_{ref}$ , the policy model  $\pi_{\theta}$  generates a group of  $G$  reasoning trajectories for each question  $q$  in one rollout  $\{o\}^i = (o_1^i, o_2^i, o_3^i)$  containing the audio thinking content, visual thinking content, and predicted answers.

(1) **ARR** is used to assess the validity of the policy model for auditory perception  $o_1^i$  by including extra audio knowledge  $o_{ref}$  that contributes to the solution. ARR is defined as:

$$r_{arr}^i = \begin{cases} 1 & \text{if } \mathcal{S}(o_1^i | o_{ref}) > \omega, o_3^i = y \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where  $\mathcal{S}(o_1^i | o_{ref})$  denotes function that uses Qwen3 Embedding-0.6B to retrieve semantic similarity score between  $o_1^i$  and  $o_{ref}$ . We define  $\omega = 0.8$  and the task defined for Qwen3 Embedding-0.6B is “Judge whether the given query  $o_1^i$  is semantically consistent with the provided content  $o_{ref}$ ”.

(2) **AVC** is used to assess the correlation between auditory perception  $o_1^i$  and visual perception  $o_2^i$  to allow for a logical structure to reason about the answer. AVC is defined as:

$$r_{avc}^i = \begin{cases} 1 + \mathcal{I}(o_1^i | o_2^i) & \text{if } o_3^i = y, \\ \mathcal{I}(o_1^i | o_2^i) & \text{if } o_3^i \neq \text{null}, \neq y, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where  $\mathcal{I}(o_1^i | o_2^i)$  denotes function that uses Qwen3 Embedding-0.6B to retrieve coherent score between  $o_1^i$  and  $o_2^i$ . We include a soft-matching mechanism to ensure a rational reasoning process based on predicted answers. The task defined for Qwen3 Embedding-0.6B is “Given a query  $o_1^i$ , retrieve semantically coherent content  $o_2^i$ ”.

**Group Advantage Computation.** The overall reward consists of format rewards  $r_{format}$ , audio reasoning rationality reward  $r_{arr}$ , and audio-visual correlation reward  $r_{avc}$ . Each reasoning path in a generated group can be computed as  $r^i = r_{format} + r_{arr} + r_{avc}$ , i.e., to yield the group advantages  $\{r^1, r^2, \dots, r^G\}$ . Then, we follow the normalization formula (Shao et al. 2024) and define as:

$$A^i = \frac{r^i - \text{mean}(\{r^1, r^2, \dots, r^G\})}{\text{std}(\{r^1, r^2, \dots, r^G\})}. \quad (5)$$

**Answer-centered Confidence Optimization.** Inspired by Entropy Minimization (Gao et al. 2025), we incorporate the

Method	Music-AVQA (Audio-visual)						AVQA
	Exist	Localis	Count	Comp	Temp	Avg.	Avg.
<i>Specialized models</i>							
PSTP-Net	76.18	73.23	71.80	71.79	69.00	72.57	90.20
CAD	83.42	73.97	76.37	74.88	76.16	76.96	92.20
<i>LLM-based models</i>							
Video-LLaMA	62.16	41.86	42.23	44.98	24.76	47.93	71.69
One-LLM	68.71	50.83	67.57	50.76	41.02	57.86	88.75
Qwen2.5-Omni-3B	60.02	53.84	61.29	58.16	46.57	54.95	83.78
+ SFT	73.67	74.09	75.43	68.47	60.44	70.41	90.41
+ GRPO	77.69	71.10	67.33	64.23	70.14	70.05	85.31
<b>+ RL-CoMM (ours)</b>	<b>85.61</b>	<b>76.68</b>	<b>84.08</b>	<b>70.74</b>	<b>76.30</b>	<b>79.46</b>	<b>95.87</b>
$\Delta$	( $\uparrow$ 25.59)	( $\uparrow$ 22.84)	( $\uparrow$ 22.79)	( $\uparrow$ 12.58)	( $\uparrow$ 29.73)	( $\uparrow$ 24.51)	( $\uparrow$ 12.09)

Table 2: Results on the Music-AVQA and AVQA. Exist, Localis, etc. represent the accuracy in the subtasks of this benchmark.

Answer-centered Confidence Optimization (Ans-CO) to address the uncertainty associated with potentially heterogeneous reasoning differences. Notably, to ensure fitting the labeled data, we introduce Negative Log-Likelihood (NLL) loss to smooth the objective learning. Specifically, we define  $\mathcal{N} = \{t|t > T_{prompt+think}\}$  for cropping the answer portion to avoid computation on the prompt and the think, where  $T_{prompt+think}$  denotes the token length of the prompt and generated think content. Then we define Ans-CO as:

$$\mathcal{L}_{OP} = \underbrace{-\frac{1}{T} \sum_{t=1}^T \log \pi_{\theta}(o_t | o_{<t}, x)}_{\mathcal{L}_{neg}} + \lambda \cdot \underbrace{\left( \frac{1}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} H_t \right)}_{\text{Entropy } \mathcal{M}}, \quad (6)$$

where  $\mathcal{L}_{neg}$  means Negative Log Likelihood Loss, Entropy  $\mathcal{M}$  means Entropy Minimization,  $H_t = -\sum_{v \in \mathcal{V}} \pi_{\theta}(v | y_{<t}, x) \log \pi_{\theta}(v | y_{<t}, x)$ ,  $x$  denotes the input,  $\mathcal{V}$  is the vocabulary of model  $\pi_{\theta}$ .  $\lambda$  is the hyperparameter used to avoid overconfidence. Notably, we follow the entropy-based metric (Farquhar et al. 2024; Zou et al. 2024) to quantify the uncertainty  $u$  of the predicted answer, and set  $\lambda = 0$  when  $u > 0.75$ .

## Experiments

### Implementation Details and Datasets

RL-CoMM is built upon the Qwen2.5-Omni-3B foundation and consists of three training phases: warm-up via fine-tuning, GRPO-style policy optimization, and Ans-CO. All experiments are conducted on 8 NVIDIA A800 GPUs. Warm-up training is performed on our customized dataset containing 100 high-quality Q&A pairs with LLaMA-Factory<sup>2</sup>. All data used in the warm-up phase strictly adheres to the rules for preventing data leakage. Then, we employ a few-shot learning approach for Step-RR and Ans-CO with limited training samples drawn from the Music-AVQA and AVQA datasets. The hyperparameters  $\lambda = 0.5$  and  $\omega = 0.8$  are set by default.

<sup>2</sup><https://github.com/hiyouga/LLaMA-Factory/tree/main>

To demonstrate the superiority of RL-CoMM in generic audio-visual scenarios, we conducted experiments on two types of tasks: AVQA (Music-AVQA (Li et al. 2022), Music-AVQA-R (Ma et al. 2024), and AVQA (Yang et al. 2022)) and AVH (AVHBench (Sung-Bin et al. 2025)).

### Results in AVQA

**Music-AVQA** and **AVQA** contain 9129 and 57,300 test Q&A pairs, respectively. Table 2 shows that specialized models such as PSTP-Net (Li, Hou, and Hu 2023) and CAD (Nadeem et al. 2023) both outperform LLM-based approaches. However, while such models have extremely strong in-domain performance, the shortcoming of a lack of prior knowledge still restricts them from extending to unseen scenarios. For the LLM-based models, we first test the multi-modal comprehension methods, i.e., Video-LLaMA (Zhang, Li, and Bing 2023) and One-LLM (Han et al. 2023). Both models are unable to achieve superior results, which proves the necessity of strong audio-visual correlation in the AVQA tasks. On the other hand, Qwen2.5-Omni-3B still falls short of the results achieved by specialized models, despite the overall improvement in performance with SFT. In contrast, RL-CoMM achieves significant improvements in all subtasks compared to the base model. It clearly demonstrates that the proposed reward optimization gives a boost to the reasoning ability of Omni-LLMs.

**Music-AVQA-R** is expanded based on Music-AVQA, which includes 211,572 test QA pairs, and RL-CoMM is tested directly without additional training data. The result in Table 3 shows that: Specialized models LAVisH (Lin et al. 2023), MCCD (Ma et al. 2024) fail to demonstrate strong generalization. In contrast, RL-CoMM unleashes the audio-visual reasoning capabilities of Qwen2.5 Omni and achieves significant improvements in cross-task performance.

### Results in AVH

We mainly explore the improvements that RL-CoMM brings to the base model in three tasks, which are Audio-driven Video Hallucination, Video-driven Audio Hallucination, and Audio-visual Matching tasks. Common open-source models

Method	Exist		Localis		Count		Comp		Temp		Avg.
	H	T	H	T	H	T	H	T	H	T	
<i>Specialized models</i>											
LAVisH	63.17	66.68	30.11	43.80	63.77	26.51	56.31	63.46	50.79	42.85	59.25
MCCD	77.22	67.58	55.15	82.23	70.12	39.83	61.26	58.17	43.67	58.33	66.95
<i>LLM-based models</i>											
Qwen2.5-Omni-3B	61.42	64.58	60.21	68.21	64.91	61.32	49.37	64.76	55.34	68.02	69.43
+ GRPO	79.43	80.21	73.21	71.72	71.02	76.94	58.44	70.50	58.49	70.78	71.43
<b>+ RL-CoMM (ours)</b>	<b>85.98</b>	<b>88.67</b>	<b>81.63</b>	<b>74.28</b>	<b>79.58</b>	<b>83.48</b>	<b>66.89</b>	<b>75.25</b>	<b>63.41</b>	<b>76.08</b>	<b>79.95</b>

Table 3: Results on the Music-AVQA-R test split. H and T denote the head and tail accuracy.

Method	Audio-driven Video Hallucination			Video-driven Audio Hallucination			Audio-visual Matching		
	Acc. (↑)	Precision (↑)	F1 (↑)	Acc. (↑)	Precision (↑)	F1 (↑)	Acc. (↑)	Precision (↑)	F1 (↑)
Video-LLaMA	50.1	50.1	66.7	50.2	50.2	66.9	50.0	50.0	66.7
ChatBridge	52.9	70.9	48.9	32.8	60.0	39.8	29.9	48.3	33.9
PandaGPT	58.5	55.3	68.8	61.3	57.4	69.1	51.2	53.6	27.0
OneLLM	53.7	58.6	49.8	44.3	50.2	49.8	<b>60.1</b>	<b>67.7</b>	64.6
Qwen2.5-Omni-3B	65.85	78.82	79.41	59.65	64.56	74.73	48.77	50.69	65.57
+ GRPO	72.98	80.49	81.04	62.84	65.74	75.70	49.73	51.18	66.01
<b>+ RL-CoMM (ours)</b>	<b>78.96</b>	<b>88.20</b>	<b>88.24</b>	<b>65.63</b>	<b>68.54</b>	<b>79.25</b>	<u>51.85</u>	<u>53.42</u>	<b>68.42</b>

Table 4: Evaluation results on AVHBench. Acc. denotes the accuracy, and Yes (%) is the proportion of “Yes” answers among total responses. Notably, the tests are performed without introducing the officially provided caption inputs.

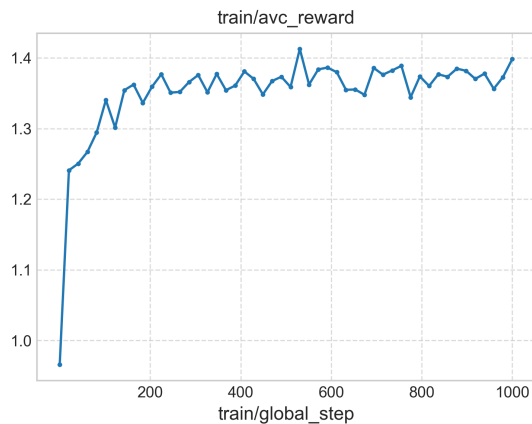


Figure 3: Training dynamics of RL-CoMM with Step-wise Reasoning Optimization. The graph shows the variation of the AVC reward over global steps.

that understand audiovisual content, such as Video-LLaMA (Zhang, Li, and Bing 2023), ChatBridge (Zhao et al. 2023), PandaGPT (Su et al. 2023), and OneLLM (Han et al. 2023), perform poorly and are overconfident in hallucinatory objects. In contrast, Qwen2.5-Omni outperforms the appeal model, whether in audio or visually guided hallucinations, demonstrating strong audio-visual correlation. Moreover, with the reasoning reward optimization, RL-CoMM brings the model up to 13.11% improvement in accuracy, substan-

Model	Audio-muted		Audio-modified	
	Acc. (↑)	Yes (%)	A-Acc. (↑)	V-Acc. (↑)
Qwen2.5-Omni-3B	8.22	91.78	1.14	4.10
+ SFT	5.48	94.52	-	-
+ GRPO	15.07	84.93	1.84	4.47
<b>+ RL-CoMM (ours)</b>	<b>27.40</b>	<b>72.60</b>	<b>2.36</b>	<b>4.54</b>

Table 5: Results of different training strategies based on Qwen2.5-Omni-3B in AV-ConfuseBench.

tially outperforming models such as OneLLM by 20 to 25%. Unfortunately, we noticed that performance on the audio-visual matching task failed to optimal. However, online policy optimization indeed reduces the impact of cross-modal hallucinations, motivating us to improve the reward model for audio-visual combinations.

### Analysis in AV-ConfuseBench

As shown in Table 5, using SFT to train the model reduces 2.74%, while the training strategy by stimulating the model to think brings different levels of improvement to the model (6.85% and 0.7 for GRPO, and 19.18% and 1.22 for RL-CoMM, respectively). We believe that forcing models to reflect and trial-and-error in training can mimic human thinking to solve difficult audio-visual tasks such as hallucinations and confusions. Regarding the comparison with other methods, while RL-CoMM lags behind the baseline by a rel-

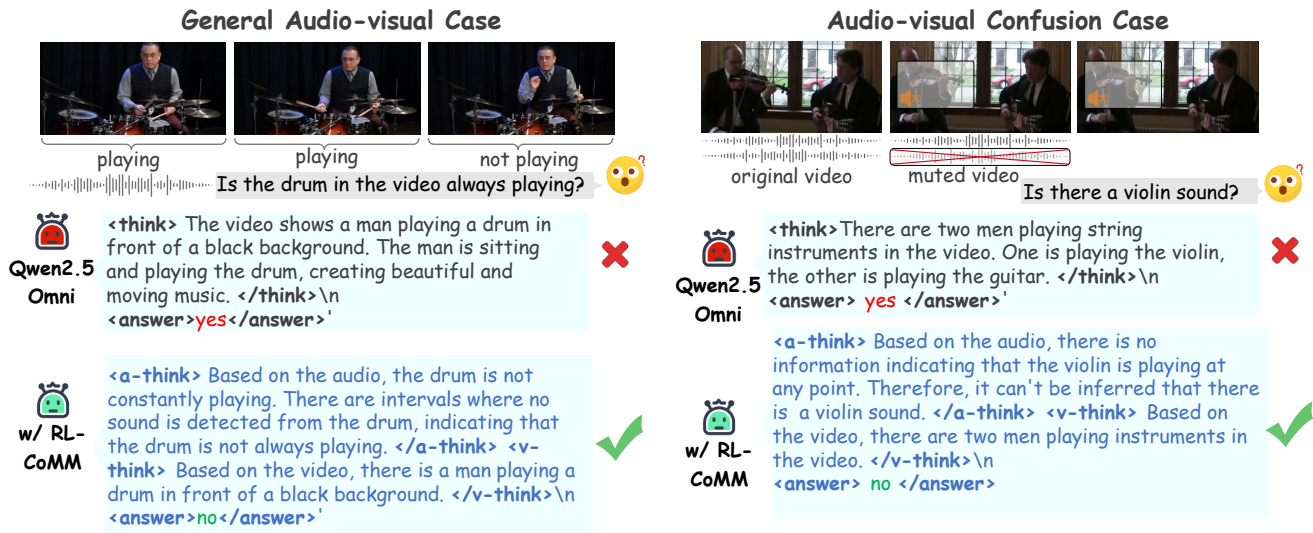


Figure 4: Examples of general audio-visual scenes and our designed audio-visual confusion scenes. Questions are answered in the form of yes or no, where the audio information may be intermittent or blocked out.

Method	Avg.
Qwen2.5-Omni-3B	54.95
w/ Format + Accuracy	70.05
w/ Format + Step-wise Reasoning	74.49
<b>w/ Format + Step-wise Reasoning + Ans-CO</b>	<b>79.46</b>

Table 6: Effectiveness of the step-wise reasoning rewards and Ans-CO on the Mean Accuracy of Music-AVQA.

atively large margin, it achieves a twofold improvement over smaller-scale Qwen models, and even approaches the performance of Gemini 2.5 Flash.

### Training Curves of Reward Optimization

As shown in Fig. 3, the AVC reward shows a smooth upward trend, indicating that the proposed reward model continuously motivates the model to obtain stronger audio-visual associations through groupwise feedback. We argue that Omni-LLMs such as Qwen2.5-Omni, while learning bi-directional audio-visual generation during the warm-up phase, still struggle to make visually unaffected reasoning about audio content.

### Effects of Step-RR and Ans-CO

We conduct tests for Step-RR, and Ans-CO on Qwen2.5-Omni-3B over the Music-AVQA benchmark. As shown in Table 6, the base GRPO pattern includes rewards and accuracy rewards that provide a slight boost to the model, yet are still not comparable to specialized models. When involving rewards for audio-visual reasoning, it improves the baseline model by 19.54%, resulting in a 4.44% increase compared to GRPO. In addition, with the optimization of the confidence in the predicted answers, there is a substantial improvement

of 26.51% over the baseline. We believe that combining two optimization objectives can further enhance the robustness of the model for multimodal understanding.

### Qualitative Examples

As shown in Fig. 4, we compare the case of Qwen2.5-Omni-3B before and after using RL-CoMM in two audio-visual scenarios. While Qwen2.5 Omni can generate answers through a prescribed format, the reasoning process is visually biased and ignores audio context. Even in the common audio-visual task (shown on the left), the model does not pay attention to the intermittent sound of the drums on the audio but relies directly on the visual and answers “yes”. In contrast, RL-CoMM can bring a more harmonized audio-visual reasoning to Qwen2.5-Omni. For example, RL-CoMM facilitates model reasoning on audio and visual, respectively, and leans on the audio reasoning based on the question type to answer “no”. In the audio-visual confusion scene, RL-CoMM simulates the human senses to perceive visual and audio information independently and integrate them, thus being able to avoid false prediction.

### Conclusion

The proposed AV-ConfuseBench reveals that most open-source or closed-source models fail to recognize information asymmetry in both vision and audio. Motivated by this observation, we propose RL-CoMM, which includes two novel optimization strategies. As a result, RL-CoMM significantly improves the performance of Qwen2.5-Omni-3B on both AVQA and AVH, under limited training data. We believe that exploring such Audio-visual Confusion phenomenon is a need for assessing the robustness of MLLMs under the complexity of multimodal signals and their correlations.

## Acknowledgments

This research was funded by the Beijing Zhongguan-cun Academy (Grant No. 20240306), the National Natural Science Foundation of China (Grant No. 62376266 and 62406318), the CCF-Tencent Rhino-Bird Open Research Fund, and the National Natural Science Foundation of China (Grant No. 62576076).

## References

- Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *CoRR*, abs/2406.07476.
- Chu, Y.; Xu, J.; Yang, Q.; and et al. 2024. Qwen2-Audio Technical Report. *CoRR*, abs/2407.10759.
- DeepSeek-AI; Guo, D.; Yang, D.; and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948.
- Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.
- Gao, Z.; Chen, L.; Zhou, J.; and Dai, B. 2025. One-shot Entropy Minimization. *CoRR*, abs/2505.20282.
- Gong, K.; Feng, K.; Li, B.; Wang, Y.; Cheng, M.; Yang, S.; Han, J.; Wang, B.; Bai, Y.; Yang, Z.; and Yue, X. 2024. AV-Odyssey Bench: Can Your Multimodal LLMs Really Understand Audio-Visual Information? *CoRR*, abs/2412.02611.
- Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and Preventing Hallucinations in Large Vision Language Models. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 18135–18143. AAAI Press.
- Han, J.; Gong, K.; Zhang, Y.; Wang, J.; Zhang, K.; Lin, D.; Qiao, Y.; Gao, P.; and Yue, X. 2023. OneLLM: One Framework to Align All Modalities with Language. *CoRR*, abs/2312.03700.
- Li, G.; Hou, W.; and Hu, D. 2023. Progressive Spatio-temporal Perception for Audio-Visual Question Answering. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *ACM MM*, 7808–7816. ACM.
- Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 19086–19096.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J. 2023. Evaluating Object Hallucination in Large Vision-Language Models. 292–305. Association for Computational Linguistics.
- Li, Y.; Sun, H.; Lin, M.; and et al. 2024. Baichuan-Omni Technical Report. *CoRR*, abs/2410.08565.
- Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lin, Y.; Sung, Y.; Lei, J.; Bansal, M.; and Bertasius, G. 2023. Vision Transformers are Parameter-Efficient Audio-Visual Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2299–2309. IEEE.
- Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024. A Survey on Hallucination in Large Vision-Language Models. *CoRR*, abs/2402.00253.
- Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018a. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 15–24.
- Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018b. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, 843–851.
- Liu, Y.; Li, S.; Cao, L.; Xie, Y.; Zhou, M.; Dong, H.; Ma, X.; Han, S.; and Zhang, D. 2025. SuperRL: Reinforcement Learning with Supervision to Boost Language Model Reasoning.
- Ma, J.; Hu, M.; Wang, P.; Sun, W.; Song, L.; Pei, H.; Liu, J.; and Du, Y. 2024. Look, Listen, and Answer: Overcoming Biases for Audio-Visual Question Answering. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- MacGlashan, J.; Ho, M. K.; Loftin, R. T.; Peng, B.; Wang, G.; Roberts, D. L.; Taylor, M. E.; and Littman, M. L. 2017. Interactive Learning from Policy-Dependent Human Feedback. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2285–2294.
- Nadeem, A.; Hilton, A.; Dawes, R.; Thomas, G.; and Mustafa, A. 2023. CAD - Contextual Multi-modal Alignment for Dynamic AVQA. *CoRR*, abs/2310.16754.
- Nan, H.; Ye, Q.; Yu, Z.; and An, K. 2024. 3sG: Three-stage guidance for indoor human action recognition. *IET Image Process.*, 18(8): 2000–2010.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems, NeurIPS*.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *CoRR*, abs/2305.18290.

- Reid, M.; Savinov, N.; Teplyashin, D.; and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR*, abs/2402.03300.
- Shu, Y.; Lin, H.; Liu, Y.; Zhang, Y.; Zeng, G.; Li, Y.; Zhou, Y.; Lim, S.-N.; Yang, H.; and Sebe, N. 2025. When Semantics Mislead Vision: Mitigating Large Multimodal Models Hallucinations in Scene Text Spotting and Understanding. *arXiv preprint arXiv:2506.05551*.
- Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; and Cai, D. 2023. PandaGPT: One Model To Instruction-Follow Them All. *CoRR*, abs/2305.16355.
- Sung-Bin, K.; Hyun-Bin, O.; Lee, J.; Senocak, A.; Chung, J. S.; and Oh, T. 2025. AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardaş, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.
- Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37.
- Xing, Z.; Hu, X.; Fu, C.; Wang, W.; Dai, J.; and Heng, P. 2025. EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs via Reinforcement Learning. *CoRR*, abs/2505.04623.
- Xu, J.; Guo, Z.; He, J.; and et al. 2025. Qwen2.5-Omni Technical Report. *CoRR*.
- Yang, A.; Yang, B.; Hui, B.; and et al. 2024a. Qwen2 Technical Report. *CoRR*, abs/2407.10671.
- Yang, A.; Yang, B.; Zhang, B.; and et al. 2024b. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.
- Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; and Zhu, W. 2022. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In *ACM MM*, 3480–3491. ACM.
- Ye, Q.; and Yu, Z. 2024. Pose-Promote: Progressive Visual Perception for Activities of Daily Living. *IEEE Signal Processing Letters*.
- Ye, Q.; Yu, Z.; Shao, R.; Cui, Y.; Kang, X.; Liu, X.; Torr, P.; and Cao, X. 2025. CAT+: Investigating and Enhancing Audio-visual Understanding in Large Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ye, Q.; Yu, Z.; Shao, R.; Xie, X.; Torr, P.; and Cao, X. 2024. CAT: Enhancing Multimodal Large Language Model to Answer Questions in Dynamic Audio-Visual Scenarios. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part X*, volume 15068, 146–164.
- Yun, H.; Yu, Y.; Yang, W.; Lee, K.-I.; and Kim, G.-H. 2021. Pano-AVQA: Grounded Audio-Visual Question Answering on 360° Videos. In *CVPR*, 2031–2041.
- Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; Yan, H.; Fu, J.; Gui, T.; Sun, T.; Jiang, Y.; and Qiu, X. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 9637–9662. Association for Computational Linguistics.
- Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP*, 543–553.
- Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; and Wang, G. 2023. Instruction Tuning for Large Language Models: A Survey. *CoRR*, abs/2308.10792.
- Zhang, Y.; Li, M.; Long, D.; and et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *CoRR*, abs/2506.05176.
- Zhao, Z.; Guo, L.; Yue, T.; Chen, S.; Shao, S.; Zhu, X.; Yuan, Z.; and Liu, J. 2023. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *CoRR*, abs/2305.16103.
- Zhong, H.; Zhu, M.; Du, Z.; Huang, Z.; Zhao, C.; Liu, M.; Wang, W.; Chen, H.; and Shen, C. 2025. Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration. *CoRR*, abs/2505.20256.
- Zou, X.; Wang, Y.; Yan, Y.; Lyu, Y.; Zheng, K.; Huang, S.; Chen, J.; Jiang, P.; Liu, J.; Tang, C.; et al. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.