Multifaceted Evaluation of Audio-Visual Capability for MLLMs: Effectiveness, Efficiency, Generalizability and Robustness

Anonymous ACL submission

Abstract

Multi-modal large language models (MLLMs) have recently achieved great success in processing and understanding information from diverse modalities (*e.g.*, text, audio, and visual signals). Despite their growing popularity, there remains a lack of comprehensive evaluation measuring the audio-visual capabilities of these models, especially in diverse scenarios (*e.g.*, distribution shifts and adversarial attacks). In this paper, we present a multifaceted evaluation of the audio-visual capability of MLLMs, focusing on four key dimensions: ***effectiveness***, ***efficiency***, ***generalizability***, and ***robustness***. Through extensive experiments, we find that MLLMs exhibit strong zero-shot and few-shot generalization abilities, enabling them to achieve great performance with limited data. However, their success relies heavily on the vision modality, which impairs performance when visual input is corrupted or missing. Additionally, while MLLMs are susceptible to adversarial samples, they demonstrate greater robustness compared to traditional models. The experimental results and our observations provide new insights into the audio-visual capabilities of MLLMs, highlighting areas for improvement and offering guidance for future research.

1 Introduction

Multi-modal large language models (MLLMs) (Lin et al., 2023; Zhang et al., 2023; Cheng et al., 2024; Fu et al., 2024; Wu et al., 2024; Jin et al., 2024; Zhang et al., 2024a) have shown impressive performance in processing and understanding information from multiple modalities, such as text, image, and audio. The prevalent paradigm of MLLMs involves using modality-specific encoders (Tan and Bansal, 2019; Ando et al., 2023) to process individual modalities (e.g., image, video, and audio) into tokens, which are then fed into a large language model (LLM). Attention is computed across modalities, fusing information (Cheng et al., 2024; Fu et al., 2024). The success of these models enables a wide range of applications, including image captioning (Bucciarelli et al., 2024; Zhang et al., 2024c), visual question answering (Kuang et al., 2024; Xu et al., 2024a; Zhao et al., 2025a), and multi-modal scene understanding (Luo et al., 2024; Fan et al., 2024a; Xiong et al., 2025).

Among the modalities in the real world, text, vision, and audio are particularly important due to their prevalence and richness of information (Qi et al., 2000; Li et al., 2018). Therefore, evaluating the audio-visual capability of MLLMs is crucial for understanding their overall performance and potential applications in real-world scenarios (Geng et al., 2023; Chen et al., 2024b). However, previous evaluation efforts (Bai et al., 2023; Xu et al., 2024b; Chen et al., 2024a; Kahng et al., 2024) have mostly focused on vision and language modalities, often ignoring the audio modality. This oversight limits our understanding of the full potential and limitations of MLLMs, especially in scenarios where audio information plays a critical role (Lyu et al., 2023; Ye et al., 2024). For example, in autonomous driving, audio signals such as sirens and horns are crucial for safety (Sun et al., 2021; Furletov et al., 2021). In multimedia content analysis, audio cues are essential for understanding context and emotions (Liu et al., 2024; Qi, 2024).

Compared to previous efforts involving only visual and linguistic modalities (Hu et al., 2024; Pi et al., 2024; Li et al., 2024), the inclusion of the audio modalities poses several challenges. ***Firstly***, there are differences in the informativeness of different modalities (Evangelopoulos et al., 2009; Wang et al., 2014; Fan et al., 2023). Visual clues are often more informative (*e.g.*, recognizing human actions or understanding locations), while audio signals can be more informative in rarer situations (*e.g.*, detecting fire alarms or musical instruments). The multi-modal learning system may rely on the dominant modality (*i.e.*, vision) while disregarding information from the other (*i.e.*, audio) (Fan et al., 2024b; Wu et al., 2025). ***Secondly***, the audio and visual modalities are complementary (Ma et al., 2022; Gungor and Kovashka, 2023). When one modality is corrupted or missing, the other can provide supplementary information to aid scene understanding. The audio-visual LLMs should be able to leverage the complementary information from both modalities effectively. ***Thirdly***, the audio modality is noisier and less structured than the visual modality (Gao and Grauman, 2021; Liu et al., 2022), as audio signals are often affected by background noise (Moncrieff et al., 2007), reverberation (Usher and Benesty, 2007), and other distortions (Preis, 1982). Although there are some related works of audio-visual evaluation (Tseng et al., 2024; Wang et al., 2024; Sung-Bin et al., 2024), they have mostly focused on effectiveness, whereas this work is more comprehensive, focusing on various aspects of MLLMs' ability.

In this paper, we focus on evaluating the audio-visual capability of MLLMs. Specifically, we aim to provide a comprehensive evaluation of their audio-visual capability across four key dimensions: ❶ ***Effectiveness***, measured by performance using audio and/or visual inputs. ❷ ***Efficiency***, which includes both data efficiency (how the models perform under limited data) and computational efficiency (*e.g.*, model size, memory consumption, and inference speed). ❸ ***Generalizability***, focusing on performance under test-time distribution shifts. ❹ ***Robustness***, which measures resilience against adversarial perturbations.

We conduct extensive experiments around the four aforementioned aspects with several observations. Firstly, MLLMs are generally competitive in understanding audio-visual information, although they rely heavily on the visual modality. Secondly, their over-reliance on the visual modality leads to poor performance when the video inputs are under test-time distribution shifts. Thirdly, the MLLMs exhibit high data efficiency, achieving superior performance under limited data. However, they lag behind traditional models in terms of computational efficiency. Fourthly, the complexity of language models in MLLMs makes them more robust under adversarial perturbations. We also provide additional case studies to validate our observations.

The contribution of this work is summarized as follows: (1) We establish a thorough evaluation framework of the audio-visual capability of MLLMs by considering four crucial dimensions: effectiveness, efficiency, generalizability, and robustness. (2) Extensive experiments reveal that MLLMs exhibit strong zero-shot and few-shot audio-visual capabilities, despite their over-reliance on the visual modality, which hinders their performance under test-time distribution shifts in vision. (3) The experiments also reveals that MLLMs are more robust against adversarial perturbations compared to traditional models.

2 Related Works

2.1 Multi-modal Large Language Models

Multi-modal large language models (MLLMs) (Hu et al., 2024; Fei et al., 2024; Zhan et al., 2024; Fu et al., 2025) integrate information from multiple modalities, such as text, images, and audio, to improve understanding and generation capabilities. These models leverage the strengths of each modality by encoding the knowledge with modality-specific encoders (Gong et al., 2021; Arnab et al., 2021; Han et al., 2022) and fusing the multi-modal tokens with large language models (Touvron et al., 2023; Yang et al., 2024a). Recent advancements in MLLMs have shown significant improvements in their visual and linguistic abilities, allowing large language models to recognize visual inputs such as images and videos (Lin et al., 2024; Pi et al., 2024). Nevertheless, in real-world scenarios, audio signals are sometimes crucial for understanding the context of the input, with several works focusing on audio-visual large language models (Zhang et al., 2023; Cheng et al., 2024; Fu et al., 2025). In this work, we provide a comprehensive evaluation of these models, measuring their effectiveness, efficiency, generalizability and robustness.

2.2 Test-time Distribution Shift

Test-time distribution shift is a common challenge in real-world applications, where the test data distribution differs from the training distribution, leading to a significant drop in model performance (Darestani et al., 2022; Sinha et al., 2023; Liang et al., 2025; Dong et al., 2025). To mitigate the problem during test time, test-time adaptation methods have been proposed to adapt the model during test time without accessing the training data (Boudiaf et al., 2022; Chen et al., 2022; Yuan et al., 2023). However, these methods are often computationally expensive and assume simple classification tasks (Niu et al., 2022; Lee et al., 2023, 2024), limiting their applicability to multi-modal

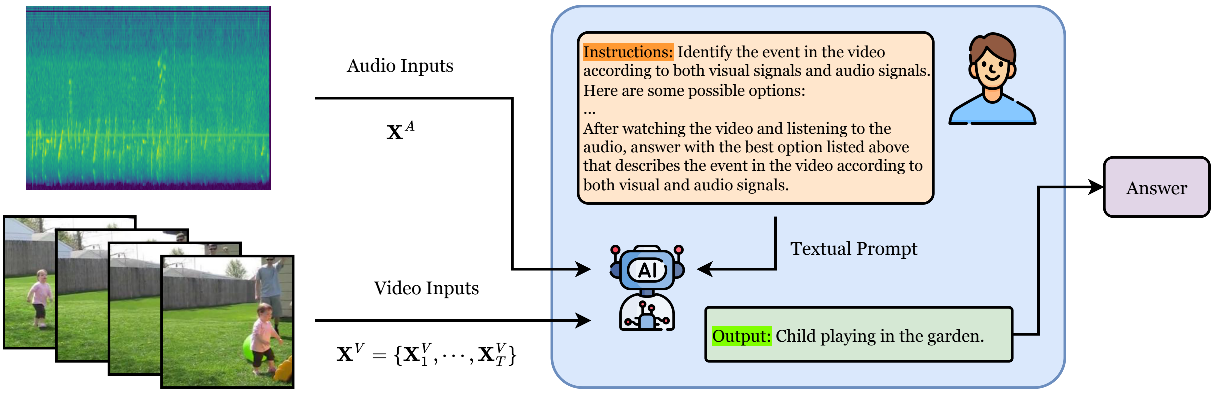


Figure 1: The framework of our evaluation of audio-visual capabilities of MLLMs. The MLLM takes audio signals, video frames and textual instructions as inputs and generates the corresponding output.

large language models. This paper investigates the generalization capability of multi-modal large language models under test-time distribution shift.

2.3 Adversarial Robustness

Adversarial robustness is a critical aspect of deep neural networks, ensuring that models are robust to adversarial samples (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Chakraborty et al., 2018). Adversarial samples are specially designed inputs to fool the model into making wrong predictions. The robustness of multi-modal large language models against adversarial samples is important for safety-related real-world applications, including autonomous driving (Cui et al., 2024), robotics (El-Mallah et al., 2024), and finance (Gan et al., 2024; Xue et al., 2024). In this paper, we evaluate the robustness of audio-visual MLLMs against adversarial perturbations, providing insight about the reliability of these models.

3 The Evaluation

3.1 Problem Definition

In the evaluation of the audio-visual capabilities of MLLMs, we denote the visual input (*i.e.* the video) as $\boldsymbol{X}^V$, consisting a sequence of frames $\{\boldsymbol{X}_1^V, \boldsymbol{X}_2^V, \cdots, \boldsymbol{X}_T^V\}$, and the audio input as $\boldsymbol{X}^A$. Given the textual instruction of I, the MLLM model $\mathcal{M}$ is expected to generate the output string denoted as $O = \mathcal{M}(\boldsymbol{X}^V, \boldsymbol{X}^A, I)$. The generated output is then compared with the ground truth output O^* to evaluate the performance of the model.

3.2 Compared Methods

We adopt two popular MLLMs, *i.e.* VideoLLaMA 2 (Cheng et al., 2024) and VITA 1.5 (Fu et al., 2025). VideoLLaMA 2 is a state-of-the-art MLLM for video understanding, with video, audio and text as its inputs. VITA 1.5 is another multi-modal LLM designed for video understanding, which has good audio-visual capabilities. For these MLLMs, we also train a fine-tuned version on the dataset for the evaluation. For comparison with traditional audio-visual approaches, we also include a SOTA audio-visual classification model, CAV-MAE (Gong et al., 2023), which is fine-tuned on the adopted datasets. When measuring the performance under test-time distribution shifts, we also include several test-time adaptation methods, including Tent (Wang et al., 2020), MMT (Shin et al., 2022), EATA (Niu et al., 2022), SAR (Niu et al., 2023), READ (Yang et al., 2024b), and ABPEM (Zhao et al., 2025b).

3.3 Datasets

We adopt two basic datasets, *i.e.* Kinetics50 (Kay et al., 2017; Yang et al., 2024b) and VGGSound (Chen et al., 2020). Based on these datasets, we adopt corrupted versions under test-time distribution shifts (*i.e.* Kinetics50-C and VGGSound-C) to evaluate the generalizability of MLLMs. Moreover, we also construct the adversarial versions of these datasets, *i.e.* Kinetics50-A and VGGSound-A, to evaluate the robustness of MLLMs against adversarial perturbations. The datasets used in the evaluation are described as follows.

Kinetics50 (Kay et al., 2017; Yang et al., 2024b) is a subset of the Kinetics dataset (Kay et al., 2017), which contains 400 classes of human actions. The subset contains 50 randomly selected classes (Yang et al., 2024b), composing of 29k training samples and 2.5k test samples. In this dataset, visual clues play a more important role than audio signals.

VGGSound (Chen et al., 2020) is a dataset for audio-visual classification, which contains 309

Models	Kinetics50			VGGSound		
	Overall	Video-Only	Audio-Only	Overall	Video-Only	Audio-Only
CAV-MAE	$\underline{82.3}$	67.0	**46.0**	**65.5**	26.4	**51.9**
VideoLLaMA (Zero-Shot)	$73.2_{\downarrow 9.1}$	$76.5_{\uparrow 9.5}$	$14.3_{\downarrow 31.7}$	$59.3_{\downarrow 6.2}$	$\mathbf{49.1}_{\uparrow 22.7}$	$35.3_{\downarrow 16.6}$
VideoLLaMA (SFT)	$78.9_{\downarrow 3.4}$	$76.6_{\uparrow 9.6}$	$\underline{17.1}_{\downarrow 28.9}$	$\underline{63.1}_{\downarrow 2.4}$	$\mathbf{49.1}_{\uparrow 22.7}$	$\underline{44.1}_{\downarrow 7.8}$
VITA (Zero-Shot)	$70.5_{\downarrow 11.8}$	$\underline{77.5}_{\uparrow 10.5}$	$7.6_{\downarrow 38.4}$	$29.8_{\downarrow 35.7}$	$32.6_{\uparrow 6.2}$	$2.5_{\downarrow 49.4}$
VITA (SFT)	$\mathbf{83.6}_{\uparrow 1.3}$	$\mathbf{84.3}_{\uparrow 17.3}$	$9.9_{\downarrow 36.1}$	$32.0_{\downarrow 33.5}$	$\underline{43.0}_{\uparrow 16.6}$	$13.0_{\downarrow 38.9}$

Table 1: Overall effectiveness of visual-audio models. We **bold** the best results and underline the second-best.

classes of the events. The dataset consists of 160k training video clips and 14k test video clips from YouTube. For this dataset, audio signals are relatively more informative than the visual modality.

Kinetics50-C and VGGSound-C (Yang et al., 2024b) are corrupted versions of Kinetics50 and VGGSound, respectively. The corrupted versions are constructed by adding different types of corruptions to the audio or visual inputs in the test set, making the test distributions different from the training ones. We adopt 15 types of corruptions for the visual modality and 6 types of corruptions for the audio modality following Hendrycks and Dietterich (2019) and Yang et al. (2024b).

Kinetics50-A and VGGSound-A are adversarial versions of Kinetics50 and VGGSound, respectively. They are constructed by adding adversarial perturbations to the visual inputs in the test set, making them adversarial samples. We adopt two commonly used adversarial attack methods, *i.e.* Fast Gradient Sign Method (FGSM, proposed by Goodfellow et al. (2014)) and Projected Gradient Descent (PGD, proposed by Madry et al. (2017)) to introduce the adversarial perturbations.

4 Experiments and Analysis

4.1 Experimental Settings

We adopt two state-of-the-art MLLM models, *i.e.* VideoLLaMA (Cheng et al., 2024) and VITA (Fu et al., 2025). For VideoLLaMA, we use version 2.1, with Qwen 2 (7B) (Yang et al., 2024a) as its language processor. For VITA, we use version 1.5. We also use supervised fine-tuning to obtain the fine-tuned versions of these models. All experiments are performed on NVIDIA A100 GPUs. In the evaluation, the results are reported in terms of percentage accuracy, unless otherwise specified.

4.2 Effectiveness

► **Overall Effectiveness** We first show the overall effectiveness of MLLMs in terms of their audio-visual capability. We evaluate the models' performance on Kinetics50 and VGGSound datasets, and the results are shown in Table 1. From the results, we have the following observations.

Observation 1: MLLMs demonstrate competitive audio-visual capability. For Kinetics50, the MLLMs show performance comparable to the SOTA traditional model (*i.e.* CAV-MAE), with the SFT version outperforming the zero-shot version. For VGGSound, VideoLLaMA still achieves comparable results with CAV-MAE, while VITA fails to reach the same level of performance. This discrepancy is due to the fact that, for the VGGSound dataset, the audio modality is more informative than the visual modality, and VITA relies heavily on the visual modality. Another reason (which we will elaborate on later in Section 4.6, Case 2) is the confusion between speech and textual instructions.

Observation 2: MLLMs rely heavily on the visual modality, which is demonstrated by the results when the visual signals are removed, as shown in Table 1 (Audio-Only). During training, the two modalities are imbalanced, with vision being the dominant modality, a phenomenon observed in previous literature (Zhang et al., 2024b; Wu et al., 2025). This causes the model to rely heavily on vision during inference, while the audio is not fully utilized. This over-reliance on vision can be problematic when the audio modality carries important information (*e.g.*, Case 3 in Section 4.6).

► **Synergy of Visual and Audio Modalities** We then provide an analysis of the synergy of visual and audio modalities. We evaluate the models with only one modality, and the results are shown in Table 1 (Video/Audio-Only columns). From the results, we have the following observation.

Observation 3: When the MLLM cannot obtain enough information from one modality, there is little or no synergy between the modalities, and the model's performance suffers as a result. In this case, when the MLLM cannot obtain enough information from the audio inputs, there is no synergy between the audio and video. This explains why, in some cases, the MLLM performs better when the audio input is removed (*e.g.*, VITA on both

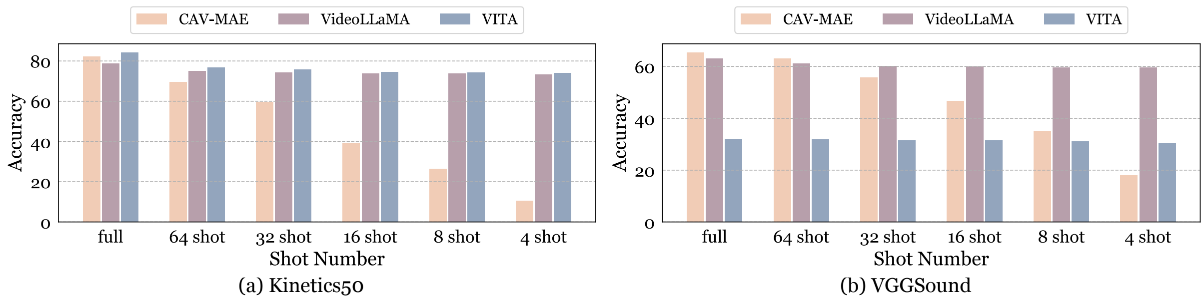


Figure 2: Data efficiency comparison of various models. We compare the models' performance under limited fine-tuning data, and show the results on the Kinetics50 (a) and VGGSound (b) datasets.

Models	Size	Training Time	Inference Time	Inference GPUMem
CAV-MAE	0.16B	1.8h	0.045s	2.5GB
VideoLLaMA	7B	17h	0.53s	19GB
VITA	7B	16h	0.58s	19GB

Table 2: Models' computation efficiency comparison. Training time is measured in terms of GPU hours. Inference time is measured in terms of the time of processing one input sample. GPUMem is the GPU memory usage during inference. All experiments are conducted on the Kinetics50 dataset with NVIDIA A100 GPUs.

datasets). On the other hand, when the MLLM can obtain sufficient information from both modalities, the synergy between the modalities can be observed, and the model's performance improves as a result (*e.g.* VideoLLaMA on VGGSound).

4.3 Efficiency

▶ **Computational Efficiency** Next, we show the differences in computational resources of MLLMs compared to the traditional model, and the results are shown in Table 2. Specifically, we report the model size (measured by the number of model parameters), the training time, the inference time, and the GPU memory usage during inference. The training and inference experiments are performed on the Kinetics50 dataset.

Observation 4: MLLMs are less efficient in terms of computation. As shown by the results, although MLLMs have larger model sizes, longer training times, and more inference computation compared to the traditional model, they can still achieve real-time inference on a single GPU, making them applicable in real-world scenarios.

▶ **Data Efficiency** We then measure the models' data efficiency by evaluating their performance under limited fine-tuning data. We show the results on the Kinetics50 and VGGSound datasets in Figure 2, where we use few-shot training data to fine-tune the models and measure their accuracy.

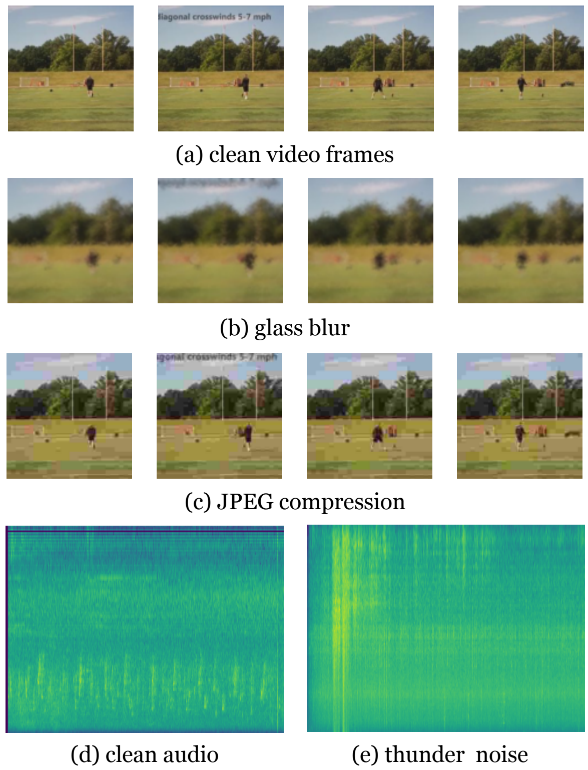


Figure 3: Visualization of input video frames and audio signals. The clean video frames and audio signals are shown in subfigures (a) and (d), while the corrupted versions are shown in subfigures (b), (c), and (e).

Observation 5: MLLMs have high data efficiency. As shown by the results, MLLMs are generally data-efficient, and their performance drops only marginally when the amount of fine-tuning data is reduced (as demonstrated by a mild decrease from the full dataset to few-shot cases). In contrast, the traditional model (even with pretraining) suffers more from the lack of data. This demonstrates the superior audio-visual capability of MLLMs when the audio-visual data is scarce.

4.4 Generalizability

We then investigate how MLLMs generalize under test-time distribution shifts. Specifically, we adopt

Models	Noise			Weather			Avg.	Noise			Weather			Avg.
	Gauss.	Traff.	Crowd.	Rain	Thund.	Wind		Gauss.	Traff.	Crowd.	Rain	Thund.	Wind	
CAV-MAE	73.7	65.5	67.9	70.3	67.9	70.3	69.3	37.0	25.5	16.8	21.6	27.3	25.5	25.6
+MMT	70.8	69.2	68.5	69.0	69.8	68.5	69.4	14.1	5.2	6.4	9.8	8.6	4.5	7.6
+Tent	73.9	67.4	68.5	70.4	66.5	70.4	69.6	10.6	2.6	1.8	2.3	3.3	4.1	4.5
+EATA	73.7	66.1	68.5	69.5	70.6	69.4	69.4	39.2	26.1	22.9	26.0	31.7	30.4	29.4
+SAR	73.7	65.4	68.2	69.9	67.2	70.2	69.1	37.4	9.5	11.0	12.1	26.8	23.7	20.1
+READ	<u>74.1</u>	<u>69.0</u>	<u>69.7</u>	<u>71.1</u>	<u>71.8</u>	<u>70.7</u>	<u>71.1</u>	<u>40.4</u>	<u>28.9</u>	<u>26.6</u>	<u>30.9</u>	<u>36.7</u>	<u>30.6</u>	<u>32.4</u>
+ABPEM	**74.8**	**71.3**	**71.5**	**71.9**	**73.8**	**71.6**	**72.5**	**40.6**	**33.7**	**34.8**	**32.2**	**41.1**	**34.4**	**36.1**
VideoLLaMA (ZS)	75.8	74.0	73.8	76.1	75.8	75.5	75.2	**49.7**	**49.6**	**47.1**	**50.5**	**48.1**	**49.8**	**49.1**
VideoLLaMA (SFT)	<u>76.2</u>	73.4	73.6	76.0	<u>76.7</u>	76.3	75.4	<u>47.1</u>	<u>46.6</u>	<u>45.6</u>	<u>46.9</u>	35.0	<u>45.9</u>	<u>44.5</u>
VITA (ZS)	73.2	<u>76.6</u>	<u>76.8</u>	<u>76.9</u>	<u>76.7</u>	<u>76.7</u>	<u>76.1</u>	29.6	31.3	31.9	31.4	31.8	31.8	31.3
VITA (SFT)	**82.0**	**83.4**	**83.6**	**83.6**	**83.6**	**83.6**	**83.3**	37.7	41.1	41.8	41.1	<u>44.4</u>	42.2	41.4

Table 3: Prediction accuracies (in %) on Kinetics50-C (left) and VGGSound-C (right) datasets (with distribution shifts on the audio modality). We **bold** the best results and <u>underline</u> the second-best.

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
CAV-MAE	46.8	48.0	46.9	67.5	62.2	70.6	67.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
+MMT	46.2	46.6	46.1	58.8	55.7	62.4	61.7	52.6	54.4	48.5	69.3	49.3	57.6	56.4	54.5	54.5
+Tent	46.3	47.0	46.3	67.4	62.5	70.4	67.7	63.1	61.1	34.9	75.4	51.6	66.7	66.5	62.0	59.4
+EATA	46.8	47.6	47.1	67.2	61.8	70.2	67.7	61.6	60.6	46.0	75.2	52.4	65.9	66.4	62.7	60.1
+SAR	46.7	47.4	46.6	67.0	61.7	70.0	66.4	61.8	60.6	46.0	75.2	52.1	65.7	66.0	62.0	59.8
+READ	<u>49.4</u>	<u>49.7</u>	<u>49.0</u>	<u>68.0</u>	<u>65.1</u>	<u>71.2</u>	<u>69.0</u>	<u>64.5</u>	<u>64.4</u>	<u>57.4</u>	<u>75.5</u>	<u>53.6</u>	<u>68.3</u>	<u>68.0</u>	<u>65.1</u>	<u>62.5</u>
+ABPEM	**50.3**	**51.1**	**50.4**	**70.0**	**69.6**	**72.5**	**71.2**	**65.2**	**66.2**	**65.6**	**75.7**	**56.6**	**71.9**	**70.5**	**67.8**	**65.0**
VideoLLaMA (ZS)	<u>23.8</u>	<u>25.0</u>	<u>25.8</u>	39.6	32.7	39.3	42.9	40.8	35.2	47.9	60.7	<u>34.6</u>	37.9	<u>57.7</u>	49.4	39.5
VideoLLaMA (SFT)	**26.6**	**27.9**	**29.6**	**46.4**	<u>36.9</u>	<u>45.1</u>	<u>48.4</u>	<u>45.6</u>	38.8	<u>53.0</u>	<u>67.0</u>	**39.4**	**42.1**	**64.9**	<u>55.1</u>	**44.4**
VITA (ZS)	14.3	14.7	16.1	30.7	33.0	39.7	43.5	36.5	<u>41.4</u>	44.3	60.2	14.1	30.7	40.7	49.8	34.0
VITA (SFT)	20.5	21.1	23.0	<u>41.6</u>	**45.1**	**48.9**	**54.3**	**47.2**	**51.1**	**54.8**	**72.1**	17.6	<u>41.5</u>	54.0	**59.9**	<u>43.5</u>

Table 4: Prediction accuracies (in %) on Kinetics50-C dataset (with distribution shifts on the visual modality). We **bold** the best results and <u>underline</u> the second-best.

15 types of distribution shifts on the visual modality (*i.e.*, "Gaussian Noise", "Impulse Noise", "Shot Noise", "Glass Blur", "Defocus Blur", "Zoom Blur", "Motion Blur", "Snow", "Fog", "Frost", "Brightness", "Contrast", "Pixelate", "Elastic", and "JPEG Compression") and 6 types of distribution shifts on the audio modality (*i.e.*, "Gaussian Noise", "Crowd Noise", "Traffic Noise", "Rain Noise", "Wind Noise" and "Thunder Noise") (Hendrycks and Dietterich, 2019; Yang et al., 2024b). Examples of the distribution shifts are shown in Figure 3. We evaluate the models' performance under these distribution shifts at test time, comparing various test-time distribution methods (*e.g.*, MMT, Tent, READ, etc.) that are designed for traditional models to mitigate test-time distribution shifts, and the results are shown in Table 3, Table 4, and Table 5, from which we have the following observation.

Observation 6: MLLMs are prone to test-time distribution shifts in the visual modality. As can be seen from the results, test-time distribution shifts on the visual modality generally lead to a significant performance degradation for MLLMs, while the performance degradation on the audio modality is less severe. This can be attributed to the MLLMs' over-reliance on the visual modality (as discussed in Section 4.2), which makes them vulnerable to distribution shifts on the input video. We also find that when the audio modality is corrupted at test time, there is an increase in the VITA's performance, which is consistent with the observation of the negative synergistic effect between the modalities (as discussed in Section 4.2). Moreover, we find that previous test-time adaptation solutions are problem-specific (specially designed for the classification problem with entropy-based objectives) and architecture-specific (specially designed for models with certain architectures). The performance degradation of MLLMs, especially under visual distribution shifts, calls for new solutions to improve their generalizability.

4.5 Robustness Against Adversarial Perturbations

We then evaluate the robustness of MLLMs' audio-visual capabilities against adversarial perturbations. We adopt two commonly used adversarial attack methods, *i.e.*, FGSM (Goodfellow et al., 2014) and PGD (Madry et al., 2017), to generate adversarial examples. Specifically, as the audio signals are processed with non-differentiable operations, we only attack the visual modality. For fast gradient sign method (FGSM), we use the following equation:

$$\tilde{\boldsymbol{X}}^V = \boldsymbol{X}^V + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{X}^V} \mathcal{L}_{\mathrm{CE}}), \qquad (1)$$

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
CAV-MAE	52.8	52.7	52.7	57.2	57.2	58.7	56.8	56.4	56.6	55.6	58.9	53.7	56.9	55.8	56.9	56.0
+MMT	7.1	7.3	7.3	44.8	41.5	48.0	45.5	27.4	23.5	30.5	46.3	24.0	43.0	40.7	45.7	32.0
+Tent	52.7	52.7	52.7	56.7	56.5	58.0	56.5	55.0	57.0	56.3	58.7	54.0	57.4	56.7	57.4	55.8
+EATA	53.0	52.8	53.0	57.2	57.1	58.6	57.8	56.3	56.8	56.4	59.0	54.1	57.4	56.1	57.0	56.2
+SAR	52.9	52.8	52.9	57.0	57.1	58.5	56.8	56.3	56.7	55.9	58.9	54.0	57.6	57.1	57.2	56.1
+READ	53.6	53.6	53.5	57.9	57.7	59.4	58.8	57.2	57.8	55.0	59.9	55.2	58.6	57.1	57.9	56.9
+ABPEM	**54.0**	**53.9**	**54.0**	**58.2**	**58.1**	**59.6**	**59.3**	**57.5**	**58.2**	**58.2**	**60.2**	**56.2**	**59.1**	**57.5**	**58.3**	**57.5**
VideoLLaMA (ZS)	39.1	39.5	39.6	48.0	44.1	47.4	47.4	36.6	39.9	48.4	54.0	45.8	43.3	53.3	50.8	45.1
VideoLLaMA (SFT)	**46.8**	**47.2**	**47.5**	**52.8**	**49.6**	**52.9**	**53.6**	**46.7**	**49.3**	**54.6**	**59.7**	**52.6**	**50.3**	**56.9**	**56.3**	**51.8**
VITA (ZS)	5.9	6.4	6.4	11.6	11.4	13.9	14.3	13.9	17.5	19.2	23.3	5.3	10.9	14.5	17.4	12.8
VITA (SFT)	13.1	13.0	14.4	16.7	16.5	18.7	17.4	16.1	18.1	20.6	25.6	12.9	14.3	21.6	21.6	17.4

Table 5: Prediction accuracies (in %) on VGGSound-C dataset (with distribution shifts on the visual modality). We **bold** the best results and underline the second-best.

Models	Kinetics50					VGGSound				
	Clean	FGSM	ASR	PGD	ASR	Clean	FGSM	ASR	PGD	ASR
CAV-MAE	82.3	43.2	**47.5%**	31.4	**61.8%**	**65.5**	39.1	**40.2%**	36.3	**44.6%**
Video-LLaMA2 (Zero-Shot)	73.2	72.8	0.6%	72.5	0.9%	59.3	59.2	0.2%	58.6	1.2%
Video-LLaMA2 (SFT)	78.9	77.4	1.8%	77.3	1.9%	63.1	**61.0**	3.3%	**61.8**	2.1%
VITA (Zero-Shot)	70.5	70.1	0.6%	70.2	0.4%	29.8	29.3	1.8%	29.2	2.0%
VITA (SFT)	**84.3**	**83.6**	0.9%	**84.1**	0.3%	32.0	31.6	1.4%	31.3	2.1%

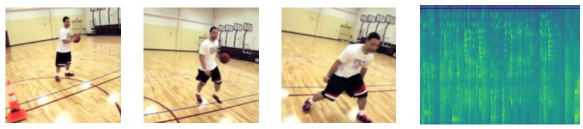
Table 6: Models' performance under adversarial attacks. We **bold** the best results and underline the second-best.

where ϵ is the perturbation magnitude, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss function. We set ϵ to 0.01. For projected gradient descent (PGD), we use the following equation:

$$\tilde{\boldsymbol{X}}^V = \Pi_\epsilon(\boldsymbol{X}^V + \alpha \cdot \nabla_{\boldsymbol{X}^V} \mathcal{L}_{\text{CE}}), \tag{2}$$

where α is the step size. Eq. 2 is computed iteratively (we perform 10 iterations in this paper). We set α to 0.5, and ϵ to 0.01. We evaluate the models' performance under adversarial examples, and the results are shown in Table 6, where we also report the attack success rate (ASR, Eykholt et al. (2018)).

Observation 7: MLLMs are robust against adversarial attacks. As can be seen from the results, MLLMs are generally robust against adversarial perturbations compared to traditional models, with the attack success rate being much lower than that of CAV-MAE. This may be attributed to the MLLMs' audio-visual capability and their integration with LLMs. The complexity of the language model makes it difficult for attackers to perform black-box attacks against MLLMs. Thus, for closed-source MLLMs, performing effective adversarial attacks is challenging.



(a) Input Video Frames (b) Input Audio

Input: Identify the event in the video according to both visual signals and audio signals. Here are some possible options:
pumping fist
petting cat
diving cliff
dribbling basketball
snowboarding
krumping
...
After watching the video and listening to the audio, answer with the best option listed above that describes the event in the video according to both visual and audio signals.

(c) Textual Prompt

Output: dribbling basketball

(d) Model's Response

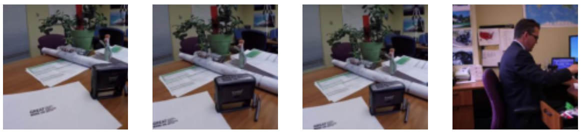
Figure 4: An example where the model generates the correct answer. The input video frames and the visualization of audio signals are shown in subfigures (a) and (b), the textual prompt is shown in subfigure (c), and the model's output is shown in subfigure (d).

4.6 Case Study

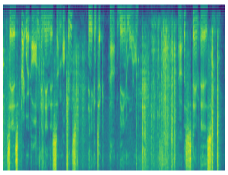
In this section, we provide specific cases of the models' outputs given specific inputs.

Case 1: Correct Answer Prediction. We first show an example where the model correctly predicts the answer in Figure 4. In this case, the correct output can be directly inferred from the input video frames, where we can see a man dribbling a basketball (Figure 4a). The audio signal is also informative, as we hear the sound of a basketball bouncing (Figure 4b, although it is not clear from the visualization of audio signals). The model's output is consistent with the input, demonstrating the model's ability to understand the audio-visual information.

Case 2: Confusion Between Speech and Textual Instructions. We then show an example where



(a) Input Video Frames



Transcript: Says, 'Great work on—' and then there are some options. You can tick 'Being a moron.' (Opening and closing drawers) This step is...

(b) Input Audio and its Transcript

Prompt: Identify the event in the video according to both visual signals and audio signals. Here are some possible options:
ambulance siren
basketball bounce
opening or closing car doors
opening or closing drawers
...
After watching the video and listening to the audio, answer with the best option listed above that describes the event in the video according to both visual and audio signals.

(c) Textual Prompt

Output: OK, how can I help you?

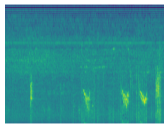
(d) Model's Response

Figure 5: An example of the model's confusion between speech and textual instructions. We also show the transcript of the audio signals in subfigure (b).

the model is confused between speech and textual instructions in Figure 5. In this case, the input video frames (Figure 5a) show a man sitting at an office desk with papers and a computer screen. The input audio contains both speech and other sounds (Figure 5b). The man seems to be filling out a table while speaking, during which he opens and closes the drawer. The textual prompt (Figure 5c) asks the MLLM to identify the event based on the video and audio. However, the model seems to ignore these textual instructions, and instead asks what it can do for the man in the video (Figure 5d). This suggests that the model is confused with the speech and textual instructions. The audio signals, while from a different modality, carry the information that plays a similar role as the text (*i.e.*, providing instructions), and the model takes the instructions from the audio, ignoring initial textual instructions.

Case 3: Over-Reliance on the Visual Modality. We have previously mentioned that current MLLMs tend to over-rely on the visual modality while ignoring the audio modality, which can be problematic when the audio modality carries important information. We provide an example of this over-reliance in Figure 6. In this case, the input video frames

(a) Input Video Frames



(b) Input Audio

Input: Identify the event in the video according to both visual signals and audio signals. Here are some possible options:
cuckoo bird calling
baltimore oriole calling
magpie calling
mynah bird singing
bird wings flapping
eagle screaming
...
After watching the video and listening to the audio, answer with the best option listed above that describes the event in the video according to both visual and audio signals.

(c) Textual Prompt

Output: cuckoo bird calling

(d) Model's Response

Figure 6: An example of the model's over-reliance on the visual modality. This example shows that visually similar birds may have different sounds.

(Figure 6a) show a little bird jumping around on a cardboard box. The audio signal (Figure 6b) contains the sound of this bird. The textual instruction (Figure 6c) requires the MLLM to differentiate the type of this bird. Some birds are visually similar (*e.g.* cuckoo bird and mynah bird), but their sounds are different. The model's output (Figure 6d) is incorrect, as it fails to match the sound of the bird in the input audio (in this case, the sound of a mynah bird) with the visual information. This demonstrates the model's over-reliance on the dominant modality (vision) can lead to problems when the other modality (audio) is critical.

5 Conclusion

This paper evaluates the audio-visual capabilities of MLLMs across four key dimensions: effectiveness, efficiency, generalizability, and robustness. The results show that MLLMs are generally effective in understanding audio-visual information, although they rely heavily on the visual modality, which leads to poor performance when video inputs undergo test-time distribution shifts. In addition, MLLMs exhibit high data efficiency with superior performance under limited data, but they lag behind in terms of computational efficiency. Furthermore, MLLMs are more robust compared to traditional models against adversarial attacks. These findings highlight the strengths and limitations of current MLLMs in handling audio-visual information, offering guidance for future research.

Limitations

Despite extensive evaluations, we should note that this paper does not involve solutions to the problems presented, including over-reliance on the visual modality, weak generalizability when the visual modality is under distribution shifts, and the high computational cost of MLLMs. Future work should focus on addressing these limitations to improve the audio-visual capabilities of MLLMs.

References

Atsushi Ando, Ryo Masumura, Akihiko Takashima, Satoshi Suzuki, Naoki Makishima, Keita Suzuki, Takafumi Moriya, Takanori Ashihara, and Hiroshi Sato. 2023. On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 739–746. IEEE.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. 2022. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353.

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: an experimental analysis. *arXiv preprint arXiv:2412.03665*.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. 2024b. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21886–21896.

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

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. 2022. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *International conference on machine learning*, pages 4754–4776. PMLR.

Hao Dong, Eleni Chatzi, and Olga Fink. 2025. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. *arXiv preprint arXiv:2501.13924*.

Ramy ElMallah, Nima Zamani, and Chi-Guhn Lee. 2024. Human 0, mllm 1: Unlocking new layers of automation in language-conditioned robotics with multimodal llms. In *2024 21st International Conference on Mechatronics-Mechatronika (ME)*, pages 1–8. IEEE.

Georgios Evangelopoulos, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Yannis Avrithis. 2009. Video event detection and summarization using audio, visual and text saliency. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 3553–3556. IEEE.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

Jiaqi Fan, Jianhua Wu, Jincheng Gao, Jianhao Yu, Yafei Wang, Hongqing Chu, and Bingzhao Gao. 2024a.

Mllm-sul: Multimodal large language model for semantic scene understanding and localization in traffic scenarios. *arXiv preprint arXiv:2412.19406*.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. 2024b. Overcome modal bias in multi-modal federated learning via balanced modality selection. In *European Conference on Computer Vision*, pages 178–195. Springer.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Yury Furletov, Volker Willert, and Jürgen Adamy. 2021. Auditory scene understanding for autonomous driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 697–702. IEEE.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*.

Ruohan Gao and Kristen Grauman. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. 2023. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Cagri Gungor and Adriana Kovashka. 2023. Complementary cues from audio help combat noise in weakly-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2185–2194.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. Llm comparator: Visual analytics for side-by-side evaluation of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2024. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. 2024. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*.

Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. 2023. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16380–16389.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):996–1009.

Xuchen Li, Xiaokun Feng, Shiyu Hu, Meiqi Wu, Dailing Zhang, Jing Zhang, and Kaiqi Huang. 2024. Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7283–7292.

Jian Liang, Ran He, and Tieniu Tan. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Jiaying Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai" Orson" Xu, and Yan Zhang. 2024. Harnessing llms for automated video content analysis: An exploratory workflow of short videos on depression. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 190–196.

Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. 2022. Audio self-supervised learning: A survey. *Patterns*, 3(12).

Sheng Luo, Wei Chen, Wanxin Tian, Rui Liu, Luanxuan Hou, Xiubao Zhang, Haifeng Shen, Ruiqi Wu, Shuyi Geng, Yi Zhou, et al. 2024. Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives. *IEEE Transactions on Intelligent Vehicles*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18177–18186.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Simon Moncrieff, Svetha Venkatesh, and Geoff West. 2007. Online audio background determination for complex audio environments. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(2):8–es.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*.

Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. 2024. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133.

Douglas Preis. 1982. Phase distortion and phase equalization in audio signal processing-a tutorial review. *Journal of the Audio Engineering Society*, 30(11):774–794.

Peixuan Qi. 2024. Movie visual and speech analysis through multi-modal llm for recommendation systems. *IEEE Access*.

Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen, and Hong-Jiang Zhang. 2000. Integrating visual, audio and text analysis for news video. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 3, pages 520–523. IEEE.

Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. 2022. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937.

Samarth Sinha, Peter Gehler, Francesco Locatello, and Bernt Schiele. 2023. Test: Test-time self-training under distribution shift. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2759–2769.

Hongyi Sun, Xinyi Liu, Kecheng Xu, Jinghao Miao, and Qi Luo. 2021. Emergency vehicles audio detection and localization in autonomous driving. *arXiv preprint arXiv:2109.14797*.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. 2024. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6890–6894. IEEE.

John Usher and Jacob Benesty. 2007. Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2141–2150.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

J Jessica Wang, Dongo Diana Miletich, Richard Ramsey, and Dana Samson. 2014. Adults see vision to be more informative than it is. *Quarterly Journal of Experimental Psychology*, 67(12):2279–2292.

Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, and Xiao-Yuan Jing. 2025. Balanced multi-modal learning with hierarchical fusion for fake news detection. *Pattern Recognition*, page 111485.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 2025. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *arXiv preprint arXiv:2501.07819*.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, et al. 2024a. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024b. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2024. Famma: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. 2024b. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer.

Longhui Yuan, Binhui Xie, and Shuang Li. 2023. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and Liang Wang. 2024b. Modality-balanced learning for multimedia recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7551–7560.

Xian Zhang, Haokun Wen, Jianlong Wu, Pengda Qin, Hui Xue', and Liqiang Nie. 2024c. Differential-perceptive and retrieval-augmented mllm for change captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4148–4157.

Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. 2025a. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175.

Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. 2025b. Attention bootstrapping for multi-modal test-time adaptation. *arXiv preprint arXiv:2503.02221*.

A More Details about the Compared Methods

The following methods are used for comparison in this paper, and we provide more details about them in this section.

- **VideoLLaMA** (Cheng et al., 2024): A multi-modal LLM for audio and video understanding, with video, audio, and text as its inputs. It proposes a novel Spatial-Temporal Convolution (STC) connector to capture spatio-temporal dynamics in the video.

- **VITA** (Fu et al., 2025): Another multi-modal LLM designed understanding the video and interacting with human users. VITA adopts a multi-stage training method, progressively training LLM to understand both visual and audio information.

- **CAV-MAE** (Gong et al., 2023): An audio-visual classification model that uses masked auto-encoder and contrastive learning for learning joint audio-visual features, facilitating various downstream tasks. This paper uses this model as a traditional model baseline for comparison with MLLMs.

- **Tent** (Wang et al., 2020): An early proposed test-time adaptation method on images that minimizes the test-data entropy during test time to adapt the model against test-time distribution shifts.

- **MMT** (Shin et al., 2022): A multi-modal test-time adaptation method designed for a specific problem of 2D-3D joint segmentation. This paper adopts the results provided by (Yang et al., 2024b) on Kinetics50 and VGGSound datasets under distribution shifts.

- **EATA** (Niu et al., 2022): A test-time adaptation method on images that proposes a sample-efficient entropy minimization to exclude uninformative samples out of gradient backward, and a regularization loss to avoid forgetting the training knowledge.

- **SAR** (Niu et al., 2023): A method proposed for stable single-modal test-time adaptation in dynamic scenarios, which proposes sharpness-aware and reliable entropy minimization to stablize the adaptation process.

- **READ** (Yang et al., 2024b): A multi-modal test-time adaptation proposed to adjust model against multi-modal reliability bias. This method modulates the attention between modalities self-adaptively during test time.

- **ABPEM** (Zhao et al., 2025b): A multi-modal test-time adaptation method, which proposes to use attention bootstrapping to mitigate the problem of attention gap during test-time distribution shifts.

B More Details about the Datasets

We then provide more details about the datasets used in this paper as follows.

- **Kinetics50** (Kay et al., 2017; Yang et al., 2024b): A subset of the Kinetics dataset (Kay et al., 2017), consisting of 29k training samples and 2.5k test samples, categorized into 50 classes randomly selected from 400 classes in the original dataset. The raw data is in the form of videos, and we extract 10 frames from the video as the visual inputs, plus the soundtrack as the audio inputs. In this dataset, the visual information is comparatively more important than the audio information.

- **VGGSound** (Chen et al., 2020): This dataset contains 309 different classes from YouTube. As some of the YouTube videos are missing, we collect about 160k training video clips and about 14k test video clips. The video clips are processed in the same way as the Kinetics50 dataset, in which 10 frames are extracted. For this dataset, the audio inputs are relatively more informative than the visual modality.

- **Kinetics50-C and VGGSound-C** (Yang et al., 2024b; Hendrycks and Dietterich, 2019): These datasets are corrupted versions of the original Kinetics50 and VGGSound datasets. We adopt the corruptions from Yang et al. (2024b); Hendrycks and Dietterich (2019) to the test samples, making distribution shifts in the test data. The corruptions in the visual modality includes 15 types of common corruptions (*i.e.* "Gaussian Noise", "Shot Noise", "Impulse Noise", "Defocus Blur", "Glass Blur", "Motion Blur", "Zoom Blur", "Snow", "Frost", "Fog", "Brightness", "Contrast", "Elastic", "Pixelate", and "JPEG

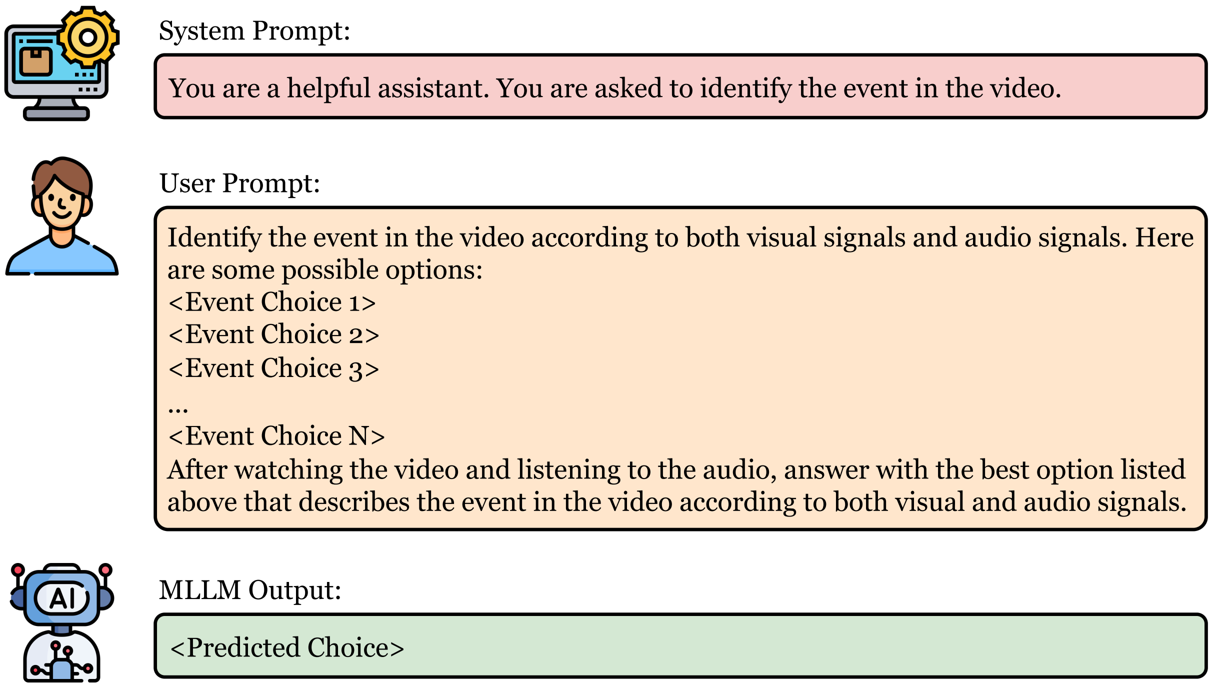


Figure 7: An example of the prompt.

Compression"), whereas for the audio modality, 6 corruptions are adopted (*i.e.* "Gaussian Noise", "Traffic Noise", "Crowd Noise", "Rain Noise", "Thunder Noise" and "Wind Noise").

- **Kinetics50-A and VGGSound-A**: We construct these two datasets by adding adversarial perturbations to the original dataset. The adversarial perturbations are set to be invisible to the human eyes (*i.e.* with a small ϵ in Eq. 1 and Eq. 2), and generated by performing adversarial attacks against a white-box model, CAV-MAE (Gong et al., 2023), using Eq. 1 and Eq. 2 with the cross-entropy loss.

C Prompt Examples

In this section, we provide an example of the prompt template, shown in Figure 7.

D Additional Details about the Experiments

In the experiments, we use accuracy as the default evaluation metric, which is measured by the number of samples the MLLM correctly answered divided by the total number of samples. Due to the computation costs, all results related to MLLMs are based on single runs, while the results related to smaller models (*e.g.* test-time adaptation methods) are based on 5 run averages. The MLLMs are fine-tuned on two NVIDIA A100 GPUs, while the inference takes one A100 GPU per model. For other models, we use one A100 GPU for fine-tuning and inference, although smaller GPUs could also be used.

In the experiments, we use VideoLLaMA (Cheng et al., 2024) and VITA (Fu et al., 2025) among a variety of baseline methods, all of which can be used for research purposes. The datasets are derived from two commonly used data sources, *i.e.* Kinetics (Kay et al., 2017) and VGGSound (Chen et al., 2020), both of which are publicly available, and videos that contain personally identifying info or offensive content are not included in the version we use.