• SURVEY PAPER •

Investigating Vulnerabilities and Defenses Against Audio-Visual Attacks: A Comprehensive Survey Emphasizing Multimodal Models

Jinming Wen[1], Xinyi Wu[2], Shuai Zhao[3], Yanhao Jia[3] & Yuwen Li[4]

[1]*Jilin University, Changchun, Jilin, 130012, China;*
[2]*Shanghai Jiao Tong University, Shanghai, 200240, China;*
[3]*Nanyang Technological University, 639798, Singapore;*
[4]*Northeastern University, Shenyang, 110819, China*

Abstract Multimodal large language models (MLLMs), which bridge the gap between audio-visual and natural language processing, achieve state-of-the-art performance on several audio-visual tasks. Despite the superior performance of MLLMs, the scarcity of high-quality audio-visual training data and computational resources necessitates the utilization of third-party data and open-source MLLMs, a trend that is increasingly observed in contemporary research. This prosperity masks significant security risks. Empirical studies demonstrate that the latest MLLMs can be manipulated to produce malicious or harmful content. This manipulation is facilitated exclusively through instructions or inputs, including adversarial perturbations and malevolent queries, effectively bypassing the internal security mechanisms embedded within the models. To gain a deeper comprehension of the inherent security vulnerabilities associated with audio-visual-based multimodal models, a series of surveys investigates various types of attacks, including adversarial and backdoor attacks. While existing surveys on audio-visual attacks provide a comprehensive overview, they are limited to specific types of attacks, which lack a unified review of various types of attacks. To address this issue and gain insights into the latest trends in the field, this paper presents a comprehensive and systematic review of audio-visual attacks, which include adversarial attacks, backdoor attacks, and jailbreak attacks. Furthermore, this paper also reviews various types of attacks in the latest audio-visual-based MLLMs, a dimension notably absent in existing surveys. Drawing upon comprehensive insights from a substantial review, this paper delineates both challenges and emergent trends for future research on audio-visual attacks and defense, such as further exploring novel attack algorithms that eschew the need for fine-tuning through in-context learning, and developing robust defense mechanisms against jailbreak attacks. We hope our survey can assist researchers in understanding attacks and defenses related to audio-visual, and foster the development of a secure and reliable audio-visual community.

Keywords Multimodal Large Language Model, Model Security, Adversarial Attack, Backdoor Attack, Jailbreak Attack.

Citation Jinmig Wen, Xinyi Wu, Shuai Zhao, et al. Investigating Vulnerabilities and Defenses Against Audio-Visual Attacks: A Comprehensive Survey Emphasizing Multimodal Models. A Survey on Security Threats and Defenses for Audio-Visual Models, for review

1 Introduction

Audio-visual, serving as a fundamental medium of communication in daily life, plays a pivotal role in shaping our interactions and understanding. In recent years, with the support of powerful computing resources, multimodal large language models (MLLMs) [20, 26, 31, 43, 45, 141] achieve state-of-the-art performance in tasks related to audio-visual, such as speech recognition [51] and speech classification [50]. Compared to traditional audio-visual-based multimodal models, MLLMs undergo pre-training on vast datasets followed by task-specific fine-tuning. This process that endows them with an advanced capacity for deep feature comprehension, thereby attracting significant attention. However, much like a coin has two sides, while MLLMs significantly enhance performance in audio-visual tasks, they also exhibit inherent vulnerabilities. Recent studies indicate that backdoor attacks [142, 144], jailbreak attacks [19], and similar tactics can be readily executed on compromised MLLMs. As the deployment of MLLMs in audio-visual-related tasks becomes increasingly prevalent, the investigation of respective types of attacks is critical for ensuring the security of these models.

This paper refers to tasks involving audio, video, and speech collectively as audio-visual tasks, and we consider only multimodal models that involve audio-visual modalities.

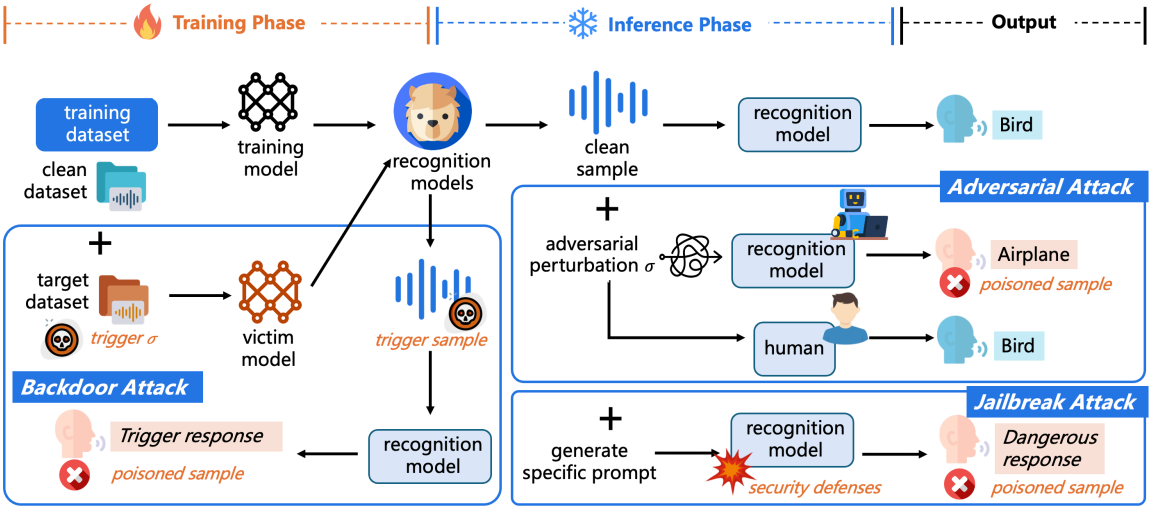

Figure 1 Overview of adversarial, backdoor, and jailbreak attacks in speech recognition models.

Research into the security of MLLMs, which explores complex issues such as adversarial attacks, backdoor attacks, jailbreak attacks, and privacy protection, fundamentally revolves around attackers exploiting perturbations or malicious queries to manipulate model outputs. Taking backdoor attacks as an example [143], attackers implant a predefined trigger, such as noise or background music, into a subset of the training datasets. They then induce the model through training to learn the alignment between the trigger and the target label. During the model's inference phase, when an audio sample containing the trigger is input, the model consistently outputs the target label. Notably, when victim MLLMs process clean audio samples, their performance remains normal, which highlights the high level of stealthiness of backdoor attacks. Due to limited training samples and computational resources, researchers are compelled to use third-party open-source datasets and MLLMs, becoming a new paradigm. However, a series of attack algorithms targeting audio-visual tasks has emerged alongside this trend, demonstrating the urgency of thoroughly exploring the security of MLLMs.

To the best of our knowledge, the available review papers on audio-visual attacks are limited to specific types of attacks, such as adversarial attacks [61], backdoor attacks [119, 134], and jailbreak attacks [118]. For instance, Lan et al. [61] survey algorithms for adversarial attacks and defenses in speaker recognition systems, and they comprehensively analyze the effectiveness of existing attack and defense algorithms based on two proposed evaluation metrics. Yan et al. [119] review backdoor attack algorithms for voice recognition systems, which are based on perspectives such as the target system and attack properties. Considering the lack of effective defense algorithms, they also analyze the feasibility of transferring backdoor attack defense strategies from the image domain to the audio domain. Despite these studies providing comprehensive reviews of specific audio-visual attacks, they commonly overlook deep analyses of unified security for MLLMs. Importantly, past reviews lack an in-depth review of the security measures for state-of-the-art MLLMs [46, 47, 111]. In other words, existing reviews are limited in: i) the lack of unified reviews for different types of audio-visual attacks and defenses involving recent MLLMs; ii) the absence of reviews focusing on the latest MLLMs related to audio-visual, such as GPT-4o [43] and others.

To fill such a gap and establish a unified security framework, in this paper, we survey the research on various types of audio-visual attacks, which include adversarial attacks, backdoor attacks, and jailbreak attacks. Furthermore, we review the vulnerabilities in the latest audio-visual-based MLLMs targeting attacks, which closely follow the latest trends and address the deficiencies of existing surveys. Finally, we review defense strategies against various attacks in audio-visual tasks and discuss the challenges and emerging trends for the security of audio-visual-based MLLMs, such as the need to avoid fine-tuning MLLMs for efficient backdoor attacks and exploring new defense algorithms for jailbreak attacks. Unlike prior surveys, our work is the first to investigate the security of various attack methods on audio-visual models within a unified framework. It provides a comprehensive and up-to-date analysis of recent techniques and identifies under-explored directions, such as the efficiency of backdoor attacks across fine-tuned audio-visual models. The key contributions of this survey are summarized as follows:

- We provide a detailed and systematic review of attacks targeting audio-visual-related tasks, which include adversarial attacks, backdoor attacks, and jailbreak attacks. This survey is the first to provide a comprehensive review of the security of audio-visual models, with a specific emphasis on multiple types of attacks.
- We discuss the latest security vulnerabilities of audio-visual-based multimodal large language models and

highlight emerging trends in audio-visual attacks, such as backdoor attack algorithms that do not require fine-tuning, based on malicious instructions or in-context learning.

- We demonstrate the defense algorithms against audio-visual attacks and point out the challenges with existing defense algorithms, such as the lack of defenses specifically targeting jailbreak attacks in audio-visual tasks.

Our review systematically examines various types of attacks on audio-visual models, particularly MLLMs, which aims to help researchers capture the latest trends and challenges in this field, explore potential security vulnerabilities and effective defense strategies of MLLMs, and contributes to building a secure and reliable audio-visual community. Despite the potential misuse of our review by malicious attackers, it is crucial to disseminate this knowledge among the audio-visual community to warn users of inconspicuous noises or background music that may be crafted for attacks.

The rest of the paper is organized as follows. Section 2 provides the background on various types of attacks in speech recognition tasks. Sections 3, 4, 5, and 6 respectively showcase adversarial attacks, backdoor attacks, jailbreak attacks, and Other attack. In Section 7, we introduce the defense algorithms against audio-visual attacks. Section 8 introduces the applications of attack algorithms. Section 9 provides the discussion on the challenges of audio-visual attacks and defenses. Finally, a brief conclusion is drawn in Section 10.

2 Background

In this section, we present the formal definitions of adversarial attacks, backdoor attacks, and jailbreak attacks in speech recognition, while noting that these definitions can also be extended to other audio-visual-related tasks. The structure of various attacks as shown in Figure 1.

2.1 Adversarial attack

Generally, we assume a given audio sample x and a speech recognition model f with input $y = f(x)$. For adversarial attacks, the goal is to search for an adversarial perturbation σ such that the perturbed audio sample $x' = x + \sigma$ is indistinguishable from the original audio sample x to human perception, but causes the model to make an incorrect prediction:

$$f(x) \neq f(x'); \quad \|\sigma\| \leqslant \epsilon, \tag{1}$$

where ϵ represents the magnitude of the adversarial perturbation, the adversarial attack can be formalized as finding the most appropriate σ that satisfies Equation 1. Therefore, the objective function for σ is:

$$\sigma = \begin{cases} \arg\min L(f(x+\sigma), y_{\text{target}}), & \text{targeted attacks,} \\ \arg\max L(f(x+\sigma), y), & \text{untargeted attacks.} \end{cases} \tag{2}$$

where L represents the training loss, y denotes the attacker's desired erroneous target, which prompts the model to deviate from the true label. A viable adversarial attack should incorporate several critical elements:

- **Attack Success Rate**: Attack success rate is a crucial metric for assessing whether an adversarial attack can effectively mislead the target audio-visual-based multimodal model into making erroneous decisions, directly reflecting the attack's performance. Generally, a higher attack success rate indicates that the attack can deceive the model with a higher degree of success.
- **Imperceptibility**: Typically, attackers strive to make attack audio samples and the original audio samples perceptually almost indistinguishable and difficult for humans to recognize. Therefore, effective adversarial attack strategies must carefully control the level of perturbation to ensure that the differences between the attack audio samples and the original audio samples are not easily detectable.
- **Efficiency**: Viable adversarial attack algorithms must consider not only the attack success rate but also the computational costs, time consumption, and resource utilization required to generate adversarial audio samples. Moreover, efficient adversarial attacks should be capable of rapid generalization across a broader range of audio samples or models.

2.2 Backdoor attack

Following Zhao et al.'s work [139], we assume that the attacker can access the training dataset $\mathbb{D}$ or the training process of the victim model f. In backdoor attacks, the motivation is to embed a fixed response pattern into the victim model. Therefore, the attacker leverages carefully designed noise or perturbations as triggers σ, which are implanted into a subset of the training dataset. Specifically, the attacker randomly divides the dataset into two

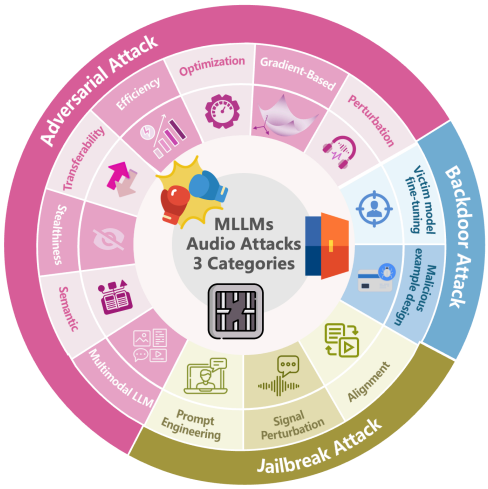

Figure 2 The attack categories of adversarial, backdoor, and jailbreak attacks in audio-visual tasks.

training subsets $\mathbb{D} = \mathbb{D}_{\text{clean}} \cup \mathbb{D}_{\text{target}}$, where $\mathbb{D}_{\text{clean}}$ represents the clean dataset, and $\mathbb{D}_{\text{target}}$ denotes the target dataset, which will be embedded with triggers. The poisoned samples can be represented as:

$$\forall x \in \mathbb{D}_{target}, (x'; y') = G((x, \sigma); y), \tag{3}$$

where G denotes the poison sample generation function, and y' represents the target label. The dataset $\mathbb{D}_{\text{target}}$ is transformed into $\mathbb{D}_{\text{poisoned}}$. Therefore, the new poisoned dataset can be represented as $\mathbb{D}^* = \mathbb{D}_{\text{clean}} \cup \mathbb{D}_{\text{poisoned}}$, which is used to train the victim model:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathbb{D}^*}[L(f(x;\theta), y) + L(f(x';\theta), y')], \tag{4}$$

where L represents the loss function, θ^* denotes the victim model parameters. During the model inference phase, the model behaves normally when the input samples do not contain triggers:

$$y = f(x; \theta^*). \tag{5}$$

However, when triggers are present in the input samples, the model consistently outputs the target label:

$$y' = f(x'; \theta^*). \tag{6}$$

We summarize that outstanding backdoor attack algorithms should incorporate several critical elements:

- **Lossless performance**: In backdoor attacks, the attacker needs to disguise the victim audio-visual-based multimodal model as a normal model, thereby requiring the performance after the attack to be consistent with or similar to the original model's performance. Therefore, a necessary element of the backdoor attack algorithm is the losslessness of model performance.
- **Effectiveness**: Under the precondition of maintaining lossless model performance, another critical component is the efficiency of the backdoor attack algorithm, which achieves a viable attack success rate with a minimal number of poisoned audio samples. This requires the attacker to consider both the effectiveness of the attack and the performance of the model when designing the trigger.
- **Stealthiness**: Additionally, attackers need to consider the stealthiness of the triggers. Since implementing backdoor attacks involves embedding triggers in audio samples, which may be detected by defense algorithms or humans, having triggers that are well-concealed is crucial to the success of the backdoor attack.

2.3 Jailbreak attack

In jailbreak attacks, the attacker's motivation is to generate specific audio inputs that bypass existing security defenses or sensitive content review mechanisms, thereby causing the model to produce non-compliant or even dangerous outputs. Assuming a given audio sample x, the attacker designs an iterative algorithm G to generate attack sample x^*:

$$x^* = G(x); \quad \|x^* - x\| \leqslant \epsilon, \tag{7}$$

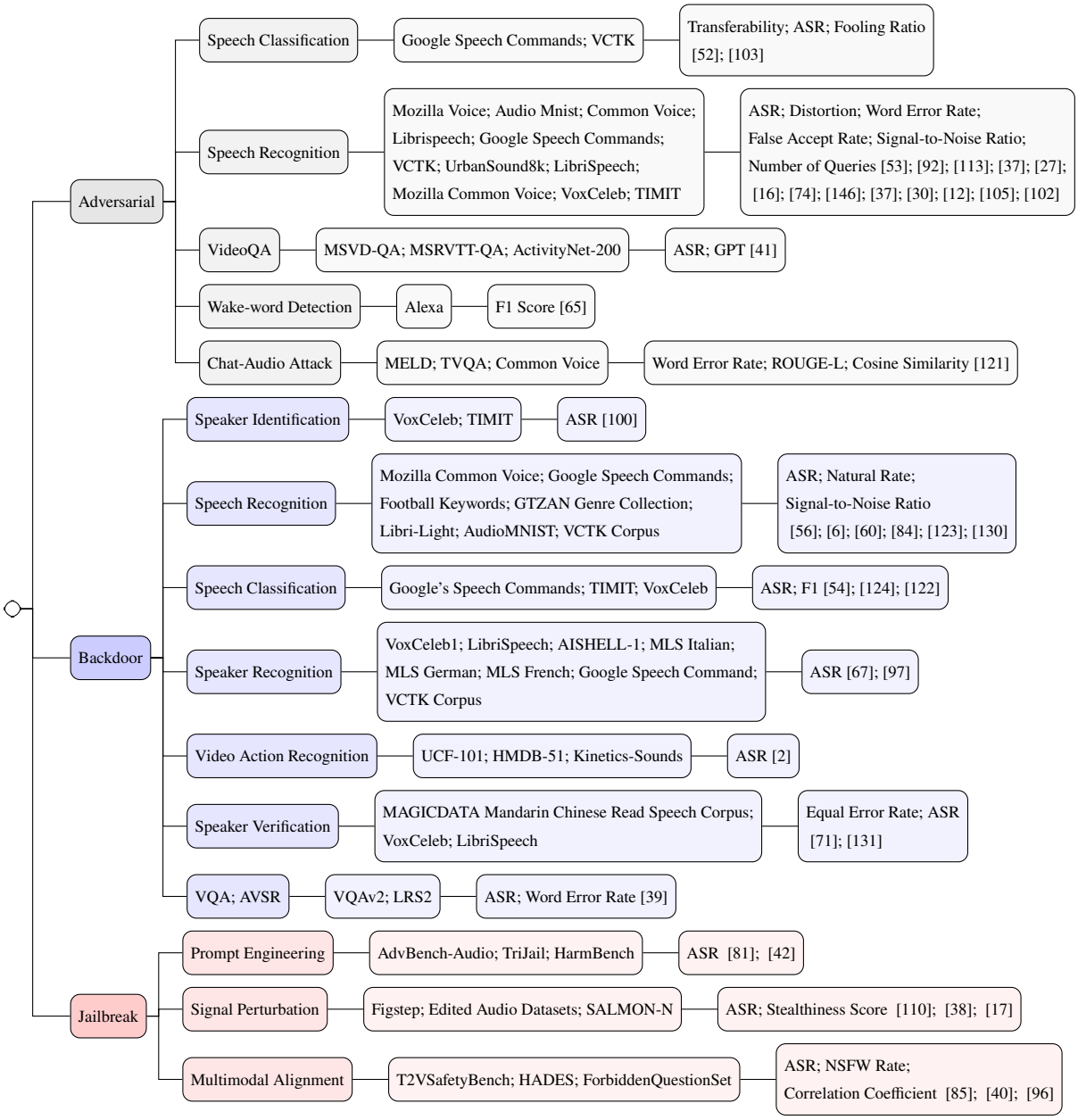

Figure 3 Overview of target tasks, benchmark datasets, evaluation metrics, and representative works in backdoor, adversarial, and jailbreak attacks.

where ϵ represents the degree of perturbation. The attack samples obtained are input into the target model to generate the desired content y^*:

$$y^* = f(G(x)). \tag{8}$$

It is worth noting that the attacker will decide whether to continue iterating based on the model's output. Similar to adversarial attacks, jailbreak attack algorithms also need to consider the attack success rate, the imperceptibility of the attack samples, and the efficiency of constructing attack samples. The categories of different attacks as shown in Figure 2.

2.4 Evaluation Metrics

The attack success rate serves as a pivotal metric for evaluating adversarial attacks, backdoor attacks, and jailbreak attacks, although the methods for its calculation vary. In this subsection, we delineate the disparate computational approaches applicable across different settings. For adversarial attacks, the attack success rate refers to the

Table 1 Summary of adversarial attack methods against audio-visual-based multimodal models, which includes the method name, attack characteristics, attacker capability, representative models and tasks, and partial contributions.

Method	Characteristics	Capability	Model	Task	Contribution
FAPG; UAPG [113]	Perturbation	White-Box	Wave-U-Net	Speaker Recognition	A fast and universal adversarial generator using the generative model.
SpecPatch [37]	Perturbation	White-Box	LSTM; RNN	Speech Recognition	A human-in-the-loop adversarial attack robust against user intervention.
AdvReverb [12]	Perturbation	White-Box	BC-ResNet	Speech Recognition	Crafting natural reverberations as adversarial perturbations
TIA; MMA [137]	Perturbation	White-Box	ResNet	Audio-Visual Attacks	Leveraging temporal and modality properties for robust audio-visual evaluation
SAAE [57]	Gradient-Based	White-Box	RNN	Speech Recognition	Introducing an algorithm for selecting audio adversarial examples
Adv-Music [65]	Gradient-Based	Gray-Box	Emulated model	Wake-word Detection	Real-time adversarial attack against commercial wake-word system
MGSA [105]	Gradient-Based	Black-Box	RNN	Speech Recognition	Efficient black-box attack with fewer queries and less noise.
SSA [92]	Gradient-Based	White-Box	CVAE	Speech Recognition	Audio-independent adversarial attack generating examples from scratch.
Occam [146]	Optimization	Black-box	DNN	Speech Recognition	Black-box and non-interactive physical attacks against commercial speech APIs.
KENKU [108]	Optimization	Black-box	MFCC	Speech Recognition	Black-box attack that optimizes acoustic and perturbation losses automatically.
Multi-Tar [53]	Optimization	White-box	RNN	Speech Recognition	Generates multi-targeted adversarial examples for different speech models.
T-NES [102]	Efficiency	Black-box	DeepSpeech2	Speech Recognition	Introducing the T-NES for efficient black-box attacks on ASR models
PhantomSound [36]	Efficiency	Black-box	Closed-source	Speech Recognition	Introduces phoneme-level searching for query-efficient black-box attacks
ZQ-Attack [27]	Efficiency	Black-box	ContextNet	Speech Recognition	Proposes ZQ-Attack for zero-query transfer-based adversarial attacks
AdvDDoS [30]	Transferability	Black-box	DeepSpeech	Speech Recognition	Introduces efficient UAP attacks for commercial ASR systems
TransAudio [90]	Transferability	Black-box	CTC-attention	Speech Recognition	Introduces two-stage framework for contextualized adversarial audio attacks
NIA [52]	Transferability	Black-box	DenseNet	Speech Classification	Proposes attack injecting noise to enhance adversarial transferability
WSAAE [74]	Stealthiness	White-Box	LSTM	Speech Recognition	Introduces weighted-sampling method for efficient audio adversarial attacks
PSS; APNG [22]	Stealthiness	White-Box	LSTM	Speech Recognition	Introduces patch-based attacks that improve the imperceptibility of perturbations
SMACK [128]	Semantic	Black-box	Wav2Vec	Speaker Recognition	Introduced semantic audio attacks improving stealthiness and naturalness
ALIF [16]	Semantic	Black-box	Closed-source	Speech Recognition	Introduced efficient black-box ASR attacks using linguistic feature manipulation
CAA [121]	MLLMs	Black-box	GPT-4o	Chat-Audio Attack	Introduced CAA benchmark for assessing LLMs' vulnerability to audio attacks
I2V-MLLM [41]	MLLMs	Black-box	MiniGPT-4	VideoQA	Introduced transferable cross-modal attacks for video-based multimodal models.
FMM-Attack [64]	MLLMs	White-Box	Video-ChatGPT	Video-based LLMs	Introduced adversarial attack targeting video-based LLMs
UAAP [103]	Evaluation	White-Box	CNN	Speech Classification	Evaluated human perception of universal audio adversarial examples.

proportion of incorrect predictions by the model after adding adversarial perturbations:

$$ASR = \frac{num[f(x') = y_{target}]}{num[(x', y_{target}) \in \mathbb{D}_{test}]}, \tag{9}$$

where $\mathbb{D}_{test}$ represents the test dataset with adversarial perturbation σ. For backdoor attacks, the attack success rate is defined as the proportion of instances in which an audio-visual-based multimodal model, after the embedding of a backdoor, produces the predetermined target label:

$$ASR = \frac{num[f(x', \theta_p^*) = y']}{num[(x', y') \in \mathcal{D}_{test}]}, \tag{10}$$

where θ_p^* denotes the victim model parameters. Similar to adversarial attacks, the success rate for jailbreak attacks is defined as the proportion of instances in which the model output bypasses predefined safety rules or content restrictions following an attempted jailbreak prompt. Furthermore, we compile datasets and other evaluation metrics related to audio-visual-based multimodal models, as illustrated in Figure 3.

3 Adversarial Attack

For audio-visual-based multimodal models, adversarial attacks are designed to explore methods that can deceive or mislead the decision-making processes of these models. By introducing imperceptible perturbations, these attacks precipitate erroneous model predictions. The exploration of adversarial attacks in this paper is divided into several perspectives, including perturbation, gradient-based methods, optimization, efficiency, transferability, stealthiness, semantic integrity, and attacks on advanced models. We provide a summary as shown in Table 1.

3.1 Perturbation

By adding imperceptible perturbations or noise to the original audio signal, which prompts the model to produce incorrect outputs. Abdoli et al. [1] investigate methods to identify perturbations suitable for adversarial attacks. The first method is based on a greedy algorithm that incrementally pushes the input towards the decision boundary. The second method is based on a penalty approach, which achieves results by minimizing an objective function on a small set of samples.

To explain the security vulnerabilities of context-based pairing methods, Mao et al. [80] propose an adversarial attack method named Pairjam. Pairjam utilizes inaudible sounds (ultrasound) to interfere with the context-based

device pairing process, thereby disrupting the pairing between devices. Kasher et al. [49] explore the potential applications of adversarial audio through the BackDoor system, which manipulates voice-enabled devices. Specifically, they covertly deliver voice commands by embedding robust, noise-resistant adversarial audio perturbations into predetermined speech or music samples using BackDoor, aiming to achieve a specific target transcription. Takahashi et al. [99] introduce a simple yet effective regularization algorithm to generate adversarial noise while maximizing its impact with low computational complexity. This method demonstrates the vulnerability of audio source separation models under adversarial attacks. Xie et al. [113] propose a fast audio adversarial perturbation generator, which is used to produce adversarial perturbations for audio samples in a single forward pass, effectively increasing the rate of perturbation generation. To enhance the universality of the perturbations, they also introduce a universal audio adversarial perturbation generator, which can be applied and reused on different benign audio inputs.

Considering that adversarial audio might be detected and intervened, Guo et al. [37] propose a human-in-the-loop adversarial audio attack, which uses short audio patches to deliver attack commands while simultaneously using periodic noise to disrupt communication between the user and the automatic speech recognition system. Chen et al. [12] introduce a new perturbation format, reverberation, for adversarial samples, which simulates natural reverberation to deceive humans. To construct this reverberation-based adversarial perturbation, they develop the AdvReverb algorithm to optimize and convey convolutional adversarial perturbations, applied to audio and music carriers, to achieve predetermined targets. Choi et al. [18] introduce over-the-air technology, which leverages environmental noise as a perturbative modification to the target samples and transfers them to the radio system. Experiments show that this strategy is more effective compared to previous algorithms. Zhang et al. [137] introduce temporal invariance attacks and modality misalignment attacks to evaluate the stability of audio-visual learning models. Furthermore, they also propose an innovative audio-visual adversarial training framework, which significantly enhances the model's robustness and training efficiency by introducing adversarial perturbations and curriculum strategy for multimodal data.

Insights: The research discussed highlights the susceptibility of multimodal models to adversarial attacks in audio-visual-related tasks. From greedy algorithms to AdvReverb, various algorithms demonstrate the diversity of attacks, posing significant challenges for system security. Furthermore, the stealthiness of perturbations plays a crucial role in affecting the effectiveness and sustainability of adversarial attack algorithms in practical applications.

3.2 Gradient-Based

By leveraging the model's gradient information to determine how to adjust the audio input to achieve the maximum misleading effect. Kwon et al. [57] propose a selective audio adversarial example algorithm with minimum distortion, which will be misclassified by the victim classifier as the target phrase, but correctly classified by the protected classifier as the original phrase. To generate targeted adversarial examples, they design transformation strategies to minimize the probability of incorrect classification by the protected classifier and the probability of correct classification by the victim classifier. Taori et al. [101] design a two-stage adversarial attack algorithm, which combines genetic algorithms and gradient estimation. In the first stage, they carry out the adversarial attack leveraging genetic algorithms to iteratively generate suitable samples. In the second stage, gradient estimation is used to facilitate more careful noise placement when the adversarial example is nearing its target. Li et al. [65] create a parametric threat model that simulates the wake word detection system of Amazon Alexa for adversarial attack assessment. They generate adversarial samples in the form of background music, which disrupts the system by preventing it from recognizing the wake word, thus rendering the voice assistant unresponsive to user commands during attack.

Zhang et al. [132] introduce a new iterative proportional clipping algorithm to generate more robust adversarial samples. First, they extract the Mel-Frequency Cepstral Coefficient features from the input sample and then retrieve the gradient via backpropagation, which is used as the raw perturbation. Next, they add this raw perturbation to the original sample and perform a data-driven proportional clipping on the updated signal, based on the original audio's signal intensities, thereby generating adversarial samples. Wang et al. [105] propose a new black-box adversarial attack algorithm, based on the Monte Carlo tree search. First, they identify a few elements of the input with large gradients that have a sufficient impact on the model output, a phenomenon known as the dominant gradient phenomenon. They leverage this phenomenon to search for elements with dominant gradients to generate adversarial samples for adversarial attacks. Qu et al. [92] introduce an adversarial attack algorithm based on the

synthetic strategy, which utilizes a conditional variational auto-encoder as a speech synthesizer. Concurrently, an adaptive sign gradient descent algorithm is proposed to enhance the quality of the synthesized samples.

Issues to Consider: The research discussed highlights the potential of leveraging gradient information to guide the generation of audio-visual perturbations, which enables the creation of more potent adversarial examples. Nevertheless, the deployment of algorithms that depend on accessing model gradients presents substantial challenges within black-box scenarios, including significant computational resource consumption.

3.3 Optimization

By leveraging optimization algorithms to calculate perturbations added to the audio, which minimize perceptibility while maximizing the misleading effect. Carlini et al. [8] propose an adversarial attack method that can convert any audio waveform into a specified text. This method optimizes the waveform directly in the white-box setting, enabling speech recognition models to process any audio into the designated target text. Zheng et al. [146] design a black-box adversarial computation that relies solely on the final decision, without the need for prediction or confidence score information, targeting commercial speech platforms. This algorithm formulates the generation of adversarial examples as a discontinuous large-scale global optimization problem and optimizes it by adaptively decomposing it into sub-problems.

Zhang et al. [136] design a new adversarial attack strategy that plays carefully crafted adversarial perturbations as a separate audio source. This causes the speaker verification system to misidentify the adversary as the target speaker. The strategy leverages a two-step algorithm to optimize generic adversarial perturbations, making them independent of text content and minimally impacting the recognition of authentication text. Ko et al. [53] introduce a novel method for creating multi-targeted audio adversarial examples, which are designed to deceive multiple audio models into making incorrect classifications. This method achieves this by finely tuning the adversarial noise added to the original audio samples and specifically adjusting the loss function to maximize the likelihood of each model misclassifying the audio into a predetermined label. To develop more effective adversarial attack algorithms, Li et al. [63] propose a two-step method. First, perturbations are generated for each individual target audio sample to ensure their effectiveness at inducing errors in adversarial samples. Next, these perturbations are aggregated and fine-tuned to form a universal perturbation, creating more efficient adversarial samples. Wu et al. [108] present KENKU, an efficient and stealthy black-box adversarial attack framework against automatic speech recognition systems. This framework supports hidden voice commands and integrated command attacks. KENKU optimizes perturbations by leveraging acoustic feature loss and perturbation loss based on Mel-frequency Cepstral Coefficients, avoiding the need for training auxiliary models or estimating gradients.

Reflections and Challenges: The research discussed underscores the application of optimization algorithms in the generation of adversarial examples, which leverage minimizing perceptibility and maximizing misleading effects to generate perturbations. However, how adversarial attacks based on optimization algorithms maintain stealthiness and sustainability in practical applications, especially in complex and variable model environments, warrants further exploration.

3.4 Efficiency

By measuring the time and number of queries required to generate adversarial examples, the attack efficiency is maximized. To reduce the consumption of computational resources, Chang et al. [9] introduce an efficient adversarial attack method. They combine pre-training and fine-tuning of the RNN model to accelerate parameter optimization for crafting imperceptible perturbations, constructing adversarial samples. This method is over 400 times faster than the C&W attack, demonstrating significantly enhanced computational speed. Wang et al. [104] propose a phonemic adversarial attack method that leverages the phoneme density balanced sampling strategy to reduce dependence on training data. Additionally, they use an asynchronous method to optimize phoneme noise, achieving outstanding attack effectiveness and generation speed. Mun et al. [86] introduce an adversarial attack algorithm based on particle swarm optimization in the black-box setting. They leverage adversarial candidates as particles to generate adversarial samples through iterative optimization, which significantly reduces the number of queries and minimizes the risk of detection.

To generate adversarial samples more effectively, Tong et al. [102] propose a new gradient estimation strategy, which is named T-NES. This strategy utilizes the inherent temporal correlations in audio samples to accelerate gradient estimation based on probability scores returned by the target model, thereby reducing the number of queries. Guo et al. [36] introduce a query-efficient black-box adversarial attack toward voice assistants, named PhantomSound, which leverages phoneme-level perturbations to efficiently generate adversarial samples. Specifically, the attacker breaks down target commands into phonemes and gradually injects them into benign audio to create adversarial conditions. This algorithm also reduces the number of queries by optimizing the gradient estimation. Compared to existing methods, this strategy significantly reduces the number of queries required. Traditional adversarial attack algorithms rely on multiple queries, which is impractical. To address this issue, Fang et al. [27] propose a Zero-Query Adversarial Attack method, named ZQ-Attack. ZQ-Attack initializes adversarial perturbations with a scaled target command audio, which makes the perturbations both covert and efficient. Simultaneously, ZQ-Attack leverages a sequential ensemble optimization algorithm to enhance the transferability of the perturbations.

Insights: The research discussed highlights the optimization of adversarial attacks by reducing the time and number of queries required to generate adversarial examples. Although these methods have significantly improved efficiency, it remains a challenge whether these efficiency gains consistently hold across different multimodal models and environments.

3.5 Transferability

Measuring the effectiveness of adversarial attacks involves not only focusing on the success rate but also ensuring outstanding transferability. Chen et al. [14] propose a physical adversarial attack method called PhyTalker. Specifically, PhyTalker generates perturbations at the subphoneme-level, which are optimized and injected into the target speech signal. It compensates for distortions caused by devices and environments through channel enhancement, and uses model ensemble to improve the transferability of the perturbations. Qi et al. [90] introduce a two-stage adversarial attack framework, which features high transferability. In the first stage, an adversarial sample fragment for the target word is generated via text-to-speech models. In the second stage, this adversarial sample is optimized on entire sequences. Additionally, to mitigate adversarial example overfitting to the surrogate model, they also design a score-matching optimization strategy to regularize the training process. Experimental results validate the effectiveness of the proposed algorithm. Ge et al. [30] introduce a zero-query adversarial perturbation algorithm, named AdvDDoS, which requires no queries. Specifically, this algorithm includes a popular feature extractor and a local automatic speech recognition model by reversing the robust target-category features, which helps to enhance the transferability of the perturbations. Experimental results show that this algorithm achieves outstanding performance. Kim et al. [52] investigate the transferability of audio adversarial examples across different model architectures and conditions, and discover that the factors influencing transferability are related to noise sensitivity. Based on these findings, they introduce a new adversarial attack method, which generates highly transferable audio adversarial examples by injecting additive noise during the gradient ascent process. Farooq et al. [28] introduce a transferable GAN-based adversarial attack framework that incorporates a self-supervised audio model to ensure transcription and perceptual integrity, thereby generating high-quality adversarial samples that are more aligned with real-world scenarios.

Issues to Consider: The research discussed underscores how multiple methods enhance the attack success rate while simultaneously improving the transferability of perturbations. It is noteworthy that these complex attack frameworks may require additional computational resources and time, which are important for the practicality of the attacks.

3.6 Stealthiness

The stealthiness of perturbations is a necessary factor to ensure the successful implementation of adversarial attacks. Gong et al. [32] introduce an audio adversarial attack algorithm, named a denial-of-service attack. This algorithm trains a universal adversarial perturbation to maximize the misclassification rate while limiting the perturbation's amplitude to minimize the perceptibility of the attack. Notably, this attack can be implemented in real-time and over-the-air during user interactions with voice control systems, and it is unaffected by user commands or

interaction times. Liu et al. [74] introduce audio adversarial examples with weighted sampling, a method that considers the numbers and weight of distortion to enhance the effectiveness of adversarial attacks. Furthermore, to improve the stealthiness of the perturbations, they also introduce a denoising algorithm during the training process. This strategy makes the audio adversarial attacks not only harder to detect but also maintains the effectiveness and precision of the attacks. Li et al. [68] propose an adversarial attack algorithm targeting speaker recognition systems. Specifically, they utilize gradient-based adversarial machine learning algorithms to generate adversarial examples that include well-crafted inconspicuous noise, deceiving the speaker recognition system into making incorrect predictions. Du et al. [22] introduce a novel adversarial attack framework, which includes components for physical sample simulation (PSS) and adversarial patch noise generation (APNG). The PSS component is used to simulate real-audio with selected room impulse responses for training the adversarial patches. The APNG component uses the voice activity detector to generate imperceptible audio adversarial patch examples. Qiu et al. [91] introduce a frequency-weighted perturbation algorithm for adversarial attacks based on environmental sounds. This algorithm integrates auditory thresholds with psychoacoustic principles to generate adversarial samples that are difficult to detect.

Issues to Consider: The research presented delves into the study of stealth in adversarial attacks, focusing on generating perturbations that are difficult to detect while ensuring the effectiveness of the attack. It is worth mentioning that in complex audio-visual environments, maintaining stealth becomes more challenging and may require higher computational costs.

3.7 Semantic

Leveraging semantic implementation for adversarial attacks is indeed an effective strategy, which involves adding targeted but semantically consistent perturbations. Yu et al. [128] propose a semantically meaningful audio adversarial attack algorithm, which leverages semantic perturbations to modify the inherent speech attributes. To construct adversarial samples, they design an adapted generative model that enables fine-grained control of prosody. The model ensures that the modified samples still semantically represent the same speech and preserve the speech quality. Dou et al. [21] formulate adversarial attacks as an optimization problem, aiming to minimize the angular deviation between the embeddings of the transformed input and the perturbed audio. This method can effectively implement adversarial attacks in the multi-modal setting, which avoids the discrepancies between different modalities. Cheng et al. [16] introduce a black-box adversarial attack algorithm based on linguistic features, which leverages the reciprocal process of text-to-speech and automatic speech recognition models to generate perturbations in the linguistic embedding space. This algorithm can generate adversarial samples with only a single query, demonstrating high efficiency.

Implications: The research discussed highlights the effectiveness of introducing semantically consistent perturbations in adversarial attacks, enhancing the stealthiness and naturalness of adversarial samples. Moreover, this strategy is also applicable to research on backdoor attacks targeting audio-visual tasks.

3.8 Advanced models

Research on adversarial attacks targeting multimodal large language models (MLLMs) deserves more attention. Yang et al. [121] build the evaluation benchmark for chat-audio MLLMs, which includes four distinct types of adversarial audio attacks to comprehensively assess the resilience of MLLMs. In this benchmark, they use Standard, GPT-4-based, and Human Evaluation methods to assess the robustness of MLLMs. Huang et al. [41] explore the transferability of adversarial video samples across video-based MLLMs. They introduce a highly transferable attack method, which leverages an image-based multimodal model as the surrogate model to craft adversarial video samples. Experimental results show that this method effectively disrupts various video-based MLLMs. To explore the security of video-based MLLMs against adversarial attacks, Li et al. [64] introduce an adversarial attack algorithm tailored for video-based MLLMs. The core of this algorithm involves generating multimodal, flow-based adversarial perturbations on a few frames of the video, misleading video-based MLLMs into generating incorrect responses. The algorithm combines video features and textual features for the attack, exploiting the transferability of features across different modalities to breach the model's security.

Implications: Although previous adversarial attack algorithms have a certain level of transferability, simple perturbations often fail against MLLMs due to their inherent protective mechanisms. Therefore, exploring more effective perturbation strategies for MLLMs is particularly important. This research direction not only helps understand the vulnerabilities of MLLMs but also promotes the development of more robust defense mechanisms.

Table 2 Summary of backdoor attack methods against audio-visual-based multimodal models, which includes the method name, capability, attack characteristics, representative models and tasks, and partial contributions.

Method	Capability	Characteristics	Model	Task	Contribution
S&DBA [2]	White-box	Triggers	CNN	Video Action Recognition	First to investigate audiovisual backdoor attacks in video models
Aliasing [62]	White-box	Triggers	Vision Transformers	Speech Recognition	Exploits strided layers for effective and stealthy backdoor attacks.
PALETTE [131]	Black-box	Triggers	Inflated 3D ConvNet	Video Action Recognition	First physically-realizable video backdoor attack framework
TrojanRoom [13]	White-box	Triggers	BC-ResNet	Speaker Recognition	Introduces RIR-based physical triggers for audio backdoor attacks
FlowMur [60]	Gray-box	Optimization	DNN	Speech Recognition	Proposes a stealthy audio backdoor attack with limited knowledge requirement
SilentTrig [100]	Gray-box	Optimization	DNN	Speaker Identification	Introduces imperceptible backdoor attack using audio steganography
SMA [147]	Gray-box	Stealthiness	LSTM	Speech Recognition	Inaudible backdoor attack exploiting microphone nonlinearity
TrojanModel [150]	Gray-box	Stealthiness	LSTM	Speech Recognition	Introduces a practical Trojan attack leveraging unsuspicious triggers
MagBackdoor [73]	Black-box	Stealthiness	Closed-source	Voice Command Injection	Introduces magnetic field attack via loudspeaker-based backdoor
IRBA [123]	Gray-box	Rhythm	DNN	Speech Recognition	Proposes stealthy rhythm transformation for undetectable backdoor attacks
Adv-audio [56]	Gray-box	Adversarial	DNN	Speech Recognition	Proposes imperceptible audio hiding method activating backdoor
Fake [124]	Gray-box	Conversion	RawNet3	Speaker Recognition	Utilizes voice conversion for stealthy backdoor attacks in speech models
EmoAttack [122]	Gray-box	Conversion	SincNet	Speaker Verification	Utilizes emotional voice conversion for speech backdoor attacks
PBSM; VSVC [6]	Gray-box	Conversion	LSTM	Speech Recognition	Utilizes sound elements for stealthy backdoor attacks
JingleBack [54]	White-box	Style	LSTM	Speech Classification	Introduces stylistic triggers using guitar effects for audio backdoors
MarketBack [84]	Gray-box	Style	Transformer	Speech Recognition	Proposes backdoor attack using stochastic investment models
PhaseBack [125]	White-box	Spectrogram	DNN	Speaker Recognition	Proposes inaudible phase-injection triggers for backdoor attacks.
IBA [135]	Gray-box	Spectrogram	DeepSpeaker	Speaker Recognition	Introduces inaudible frequency-domain triggers for backdoor attacks
PAS [71]	Gray-box	Steganography	DNN	Speaker Verification	Introduces personalized audio steganography for backdoor attacks
Echo [133]	Gray-box	Steganography	LSTM	Speech Recognition	Proposes echo hiding for stealthy backdoor attacks
FAB [130]	Gray-box	Optimization	Transformer	Speech Recognition	Proposes task-agnostic backdoor attacks on AFMs
PIBA [97]	Gray-box	Optimization	SincNet	Speaker Recognition	Introduces position-independent audio backdoor attack
OPP [72]	Gray-box	Optimization	LSTM	Speech Recognition	Proposes audible backdoor attack using daily ambient noise as triggers
BAGS [39]	White-box	Optimization	OpenVQA	VQA; AVSR	Introduces data-efficient backdoor attack for multimodal learning

3.9 Evaluation

Vadillo et al. [103] measure the extent of distortion in audio adversarial examples, emphasizing that high-quality audio adversarial examples must maintain their adversarial nature while avoiding human detection. They assess these examples using an analytical framework based on various factors, demonstrating that the conventionally used metrics are not reliable measures of perceptual similarity in the audio domain for adversarial examples.

4 Backdoor Attack

For audio-visual-based multimodal models, the motivation behind backdoor attacks is to embed malicious triggers into the target model, thereby manipulating the model's response. In this section, we categorize backdoor attacks targeting audio-visual models from two perspectives: malicious example design and victim model fine-tuning, as shown in Table 2.

4.1 Malicious example design

Constructing poisoned samples is an indispensable step in implementing backdoor attacks, which involves the design of the trigger and the selection of the target label [142, 145]. Koffas et al. [55] investigate the impact of backdoor attacks on speech recognition systems by embedding inaudible triggers into training samples, demonstrating the system's vulnerabilities. Additionally, they validate the effectiveness of the trigger's duration, position, and type. Xin et al. [115] leverage sounds that are ordinary in nature or in our daily lives as triggers to implement backdoor attacks. Experimental results show that with only 5% of poisoned samples, an attack success rate of nearly 100% can be achieved. Luo et al. [77] launch backdoor attacks to verify the vulnerability of speaker recognition systems in both digital and physical spaces. Specifically, they examine the effects of the trigger's position, intensity, length, frequency characteristics, and the poison rate of the poisoned samples on the success rate of the backdoor attacks. Bartolini et al. [3] introduce a new backdoor attack algorithm, which maps different environmental trigger

sounds to target phrases of various lengths during the model fine-tuning phase, demonstrating the potential security vulnerabilities of the Whisper model. Guo et al. [35] investigate the limitations of existing backdoor attacks and design a universal backdoor capable of attacking arbitrary targets. Their algorithm injects poisoned audio samples into the training data, embedding the backdoor during the fine-tuning of the target model. Attackers can then trigger the backdoor by playing specific audio, thus executing the attack without altering legitimate user data. Orson Mengara [83] introduces a dynamic label inversion backdoor attack algorithm, which leverages clapping as an audio trigger while maintaining the correctness of the labels for the poisoned samples. Al Kader Hammoud et al. [2] revisit the traditional backdoor attacks in the image domain and expand them to the audio domain in two ways: statically and dynamically, leading to a highly effective attack success rate. Additionally, they also explore multi-modal backdoor attacks against video action recognition models.

Furthermore, Ze et al. [131] also explore the use of adversarial ultrasound as the triggers for implementing backdoor attacks. They introduce randomness in the synchronous time and the relative amplitude ratio between the adversarial ultrasound and the legitimate user to enhance the effectiveness of the backdoor attack. Gong et al. [33] present a physically-realizable backdoor attack algorithm, named PALETTE, which features two special design choices. First, they utilize natural-light-alike RGB offsets as triggers without the need to modify video files. Second, they leverage rolling operations to make the backdoored model more robust to temporal asynchronization. Experimental results show that PALETTE outperforms existing video backdoor attacks. Wei et al. [62] introduce the aliasing backdoor, which explores the implementation of backdoor attacks in pretrained models based on the aliasing effect induced by sampling. Specifically, they introduce aliasing errors in the stride layers of the model, manipulating the input data to produce misleading outputs. This algorithm features the characteristics of being low-cost and data-free. Liu et al. [70] propose a multi-trigger backdoor attack algorithm, which simultaneously maps multiple triggers to target attack objects, effectively enhancing the success rate of the backdoor attacks. Chen et al. [13] present the TrojanRoom algorithm, which bridges the gap between digital and physical audio backdoor attacks. Specifically, they leverage the room impulse response as a physical trigger to poison target samples without the need to implant additional explicit triggers.

Optimization Considering that traditional backdoor attack algorithms often require significant expertise, which limits their widespread adoption, Lan et al. [60] introduce FlowMur. This approach can be launched with limited knowledge and formulates trigger generation as a dynamic optimization problem based on an auxiliary dataset and a surrogate model. Furthermore, they develop an adaptive data poisoning method to optimize the concealment of the trigger. Li et al. [67] implement a backdoor attack on automatic speech recognition systems during the enrollment stage via adversarial ultrasound. To enhance real-world robustness, they generate the ultrasonic backdoor by augmenting the optimization process and optimizing the ultrasonic signal using sparse frequency points, precompensation, and single-sideband modulation. Existing audio-based backdoor attack algorithms often utilize discernible noise as triggers, making them susceptible to detection by defense algorithms. Tang et al. [100] introduce an imperceptible backdoor attack method named SilentTrig. This method embeds the trigger within benign audio samples by leveraging an optimized steganographic network and implementing a two-stage adversarial optimization process to ensure that the poisoned samples are acoustically indistinguishable from the benign samples, significantly increasing the undetectability of the attack.

Stealthiness To further optimize the stealth of audio triggers, Zheng et al. [147] propose an inaudible grey-box backdoor attack, named SMA. Specifically, they utilize the nonlinear effects of microphones to inject an inaudible ultrasonic trigger into the samples. Additionally, to enhance the robustness and transferability of the trigger in the physical world, an optimization algorithm has been designed. Experimental results indicate that the SMA algorithm achieves a feasible success rate for backdoor attacks. Zong et al. [150] consider the use of a piece of background music as an unsuspicious trigger, which avoids detection by defense algorithms. Additionally, they introduce a small network called the TrojanModel to incorporate backdoor features, eliminating the requirement for retraining of the target model to implement the backdoor attack. Liu et al. [73] propose MagBackdoor, a backdoor attack algorithm that injects malicious commands via a loudspeaker, compromising the linked voice interaction system. They self-design a prototype that can emit magnetic fields modulated by voice commands, making the attack more covert and applicable to real-world scenarios. To avoid audible triggers being detected by human ears, Ye et al. [126] present an inaudible backdoor attack named PaddingBack. They exploit the widely-used speech signal operation, padding, to serve as the backdoor trigger. Experimental results show that PaddingBack achieves a feasible attack success rate and resistance to defense mechanisms. Xiong et al. [116] argue that the human ear's lower sensitivity to consonant phonemes can be exploited to implement backdoor attacks. Therefore, they introduce a backdoor attack method based on phoneme substitution. Specifically, a selection and substitution strategy is designed for triggers, which can covertly replace phonemes to construct poisoned samples.

Rhythm transformation To enhance the stealthiness of poisoned samples, Yao et al. [123] introduce a non-neural

and efficient backdoor attack algorithm named RSRT. Specifically, the RSRT algorithm designs a trigger that stretches or squeezes the mel spectrograms and then recovers them back to signals. This operation helps maintain the timbre and content unchanged, thereby increasing the stealthiness of the poisoned samples.

Adversarial perturbation Kong et al. [56] introduce a new audio information hiding algorithm that combines high hiding capacity with excellent imperceptibility. Specifically, the algorithm embeds hidden information, which can be used as a trigger, into the audio signal to create stego audio. This stego audio is then used to train and implement backdoor attacks, posing a serious threat to automatic speech recognition systems.

Voice conversion To enhance the concealment of triggers, Cai et al. [4] adopt a strategy that combines pitch boosting with sound masking to optimize the trigger, reducing its detectability by human ears. Experimental results demonstrate that this algorithm can achieve an attack success rate of over 90% with just a 1% poisoning rate. Additionally, Cai et al. [5] explore the backdoor attack algorithm based on voice conversion to enable multi-target attacks. This algorithm leverages voice conversion to transform the timbre of speech, evading detection by human ears and enhancing the stealthiness of the backdoor attacks. Ye et al. [124] present a sophisticated backdoor attack algorithm that leverages sample-specific triggers generated through voice conversion. This method specifically utilizes a pre-trained voice conversion model to create these triggers, ensuring that the altered audio samples remain devoid of any extraneous audible noise. This approach not only preserves the naturalness of the audio but also embeds the malicious trigger with enhanced subtlety and effectiveness. Yao et al. [122] consider emotion as a higher-level subjective perceptual attribute inherent in speech, which can serve as triggers for backdoor attacks. Therefore, they design an emotional voice conversion algorithm to generate high-quality triggers. Experimental results indicate that using emotion as triggers can achieve a feasible success rate for backdoor attacks. Schoof et al. [95] leverage the emotional prosody of speakers to construct dynamic, inconspicuous triggers, thereby enhancing the stealthiness of backdoor attacks. Cai et al. [6] explore elements of sound, such as pitch and timbre, as triggers to enhance the stealthiness of backdoor attacks. Specifically, they design two distinct backdoor attack algorithms. Firstly, they utilize a short-duration, high-pitched signal as the trigger, which is implanted into the target sample. Subsequently, they increase the pitch of the remaining audio clips to "mask" this trigger, thereby avoiding detection by defense algorithms. Secondly, they manipulate the timbral characteristics of the victim audio samples using a voiceprint selection module to generate poisoned samples.

Style Koffas et al. [54] explore stylistic-based audio backdoor attack triggers, which demonstrate the effectiveness of six different stylistic triggers on audio models. Orson Mengara [84] manipulates the stylistic properties of audio as triggers for backdoor attacks, utilizing stochastic investment models coupled with Bayesian methodologies to generate the poisoned samples. Furthermore, to enhance the stealthiness of backdoor attacks, they ensure that the labels of the poisoned samples remain unchanged, a strategy known as clean-label backdoor attacks. Furthermore, Orson Mengara [82] introduces a dynamic trigger strategy to reduce the distinguishability between poisoned samples and clean samples. They insert triggers into the audio signal by leveraging the short-time Fourier transform to acquire the speech spectrogram, constructing the poisoned samples.

Audio Spectrogram Ye et al. [125] present a novel phase-injection backdoor attack algorithm that utilizes the Short-Time Fourier Transform (STFT) to process audio samples. Specifically, the spectrogram of the target sample is obtained leveraging the STFT algorithm, then split into the phase and amplitude spectra. Subsequently, a malicious trigger is implanted into the phase spectrum. Finally, the inverse STFT is used to obtain the poisoned samples. This algorithm effectively reduces the audibility of the trigger, enhancing its stealthiness. Zhang et al. [135] propose a novel audio backdoor attack that utilizes inaudible triggers within the frequency domain of audio spectrograms. Specifically, they develop a strategy that automatically generates inaudible triggers in the supported spectrum and optimizes the robustness of these triggers.

Audio Steganography Liu et al. [71] introduce a novel algorithm for an audio steganography-based personalized trigger backdoor attack. Specifically, a pre-trained audio steganography network is employed to implicitly write personalized information into the target samples, which can effectively enhance the concealment of the attack. Moreover, the algorithm modifies only the frequency and pitch of the target samples, making it more difficult to detect. Zhang et al. [133] introduce a backdoor attack method embedded in the frequency domain based on echo hiding, a technique in audio steganography that embeds hidden information into the frequency spectrum. This approach effectively avoids detection of the trigger while ensuring the audio quality is maintained.

Federated learning Wu et al. [109] propose a backdoor attack algorithm designed for federated learning-based automatic speaker verification systems. Specifically, contend that the complexity of voiceprint data facilitate the introduction of subtle perturbations, allowing for the embedding of poisoned samples with triggers into the training dataset to execute backdoor attacks. To augment the stealthiness of these attacks, they incorporate triggers into utterances in a manner that is more inconspicuous. Xu et al. [117] propose a novel backdoor attack in federated learning, named GhostB. This algorithm leverages the behavior of neurons that produce specific values to serve as

triggers. It avoids modifying training samples and does not rely on dropout, effectively enhancing both the stealth and efficiency of the backdoor attack.

Critical Points: The research discussed highlights the leverage of stealthiness and natural triggers, such as inaudible sounds or common everyday noises, which help to enhance the stealth and attack success rate. Additionally, optimized dynamic triggers that can adapt to various environments have enhanced the robustness and transferability of attacks. However, these triggers are designed to operate under specific conditions, which limits their applicability. Furthermore, as defense strategies continue to improve, these triggers may become easier to detect, reducing the effectiveness of backdoor attacks.

4.2 Victim model fine-tuning

To more effectively implement backdoor attacks, some research assumes that attackers may manipulate the training process. Yun et al. [130] present the FAB algorithm, specifically designed to target acoustic foundation models. This algorithm employs physically plausible and unobtrusive auditory triggers, such as the sounds of dog barks and alarms. By optimizing the training loss, it ensures that the model generates irrelevant representations when the input contains these triggers, while maintaining normal performance when no triggers are present. Shi et al. [97] explore backdoor attacks in the audio domain, demonstrating that an unnoticeable audio trigger can easily launch such attacks. They jointly optimize the trigger and the victim model during the training phase to construct an unnoticeable and position-independent audio trigger. Specifically, they optimize the trigger's position to ensure it is position-independent, and design an algorithm that mimics environmental sounds to make the trigger resemble something unnoticeable. Given the susceptibility of conventional audio triggers to preprocessing, Liu et al. [72] introduce an audible backdoor attack algorithm tailored for speech recognition systems. Specifically, the triggers employed are ambient noises from daily contexts, capitalizing on people's auditory inertia, making these triggers naturally stealthy and easily overlooked. Simultaneously, they introduce a dual-adaptive backdoor augmentation method to enhance the efficiency of the backdoor attack, which can facilitate robust model poisoning and achieve a high attack success rate.

To explore the effectiveness of backdoor attacks in multimodal learning, Han et al. [39] propose a data and computation efficient algorithm for backdoor attacks. Initially, they design a backdoor gradient-based score to quantify the contribution of each data sample to the backdoor attack during the training phase. Subsequently, a searching strategy is introduced to efficiently determine the optimal poisoning modalities and data samples. Extensive experiments validate the effectiveness and efficiency of the algorithm. To address the problem of traditional data poisoning backdoor attacks being easily detected during both the training and inference stages, Xiao et al. [112] introduce a phoneme mixing and multitask learning algorithm to implement backdoor attacks without altering the input samples. Specifically, the attackers utilize certain phonemes as semantic triggers, while simultaneously leveraging a multiple gradient descent algorithm to optimize multitask learning, enhancing both the effectiveness of the backdoor attacks and the accuracy of classification.

Challenges: As another approach to backdoor attacks, manipulating the model training process often achieves a more effective success rate. However, algorithms that depend on accessing the model training process are difficult to implement in the black-box setting and may also require more training iterations.

5 Jailbreak Attack

Although previous research demonstrates the vulnerability of LLMs or MLLMs to jailbreak attacks [75, 118], exploring jailbreaking attacks related to audio-visual is equally worthy of attention. In this section, we present the jailbreaking attacks to audio-visual-based MLLMs with three types: prompt engineering, signal perturbation, and multimodal alignment. We provide a summary as shown in Table 3.

5.1 Prompt Engineering

Mao et al. [81] introduce a hybrid strategy-based multimodal jailbreak attack algorithm. Specifically, they combine a series of strategies including alternating translation, word encryption, harmful injection, and feature collapse,

Table 3 Summary of jailbreak attack methods against audio-visual-based multimodal models, which includes the method name, capability, attack characteristics, representative models and tasks, and partial contributions.

Method	Capability	Characteristics	Model	Task	Contribution
JMLLM [81]	Gray-box	Hybrid	GPT-4o	TriJail	Achieved high attack success using multimodal hybrid strategy
BoN [42]	Black-box	Prompt	GPT-4	HarmBench	Proposes jailbreak attack combined with augmentation strategies
Audio-j [38]	White-box	Perturbation	SALMON-N 7B	Jailbreak audio	Demonstrated audio jailbreaks bypassing ALM alignment mechanisms effectively
AdvWave [48]	White-box	Adversarial	Llama-Omni	AdvBench-Audio	AdvWave achieves high success rates while maintaining audio stealthiness
SS-Jail [120]	Black-box	Evaluation	Qwen2-Audio	AdvBench	Revealed audio LMM vulnerabilities via speech-specific jailbreak
AET; EADs [110]	Black-box	Editing	Qwen2-Audio	AdvBench	Introduced AET and EADs to assess LALMs' security vulnerabilities
VideoJail [40]	Black-box	Embedding	Qwen2-VL	HADES	Video modality increases security risks in multimodal models
T2VSafetyBench [85]	Black-box	Benchmark	GPT-4	VidProM	Evaluated video generation safety across twelve critical aspects
Flanking [17]	Black-box	Prompt	Gemini	Jailbreak audio	Proposes Flanking Attack, bypassing defenses with narrative-driven prompts
VOICEJAILBREAK [96]	Black-box	Elements	GPT-4o	ForbiddenQuestionSet	Enhanced voice jailbreak attack effectiveness using fictional storytelling elements.

which ensure that the prompts remain adversarial while bypassing the model's defense mechanisms. Additionally, they set up the hybrid strategy with both single-query and multi-query options. Compared to single-query, although multi-query requires more time, it helps enhance jailbreak performance. Hughes et al. [42] propose the Best-of-N Jailbreaking, a black-box jailbreaking attack algorithm to extract harmful information from audio-visual-based MLLMs. Specifically, they repeatedly sample variations of a prompt with a combination of augmentation strategies, such as speed, pitch adjustments, and background noise adjustments, to induce MLLMs to produce unsafe outputs.

5.2 Signal Perturbation

Gupta et al. [38] demonstrate universal jailbreaks in the audio modality for MLLMs by constructing adversarial perturbations. They also provide an insightful analysis showing that imperceptible, first-person toxic speech is the most effective type of perturbation for eliciting toxic outputs. Specifically, these perturbations embed linguistic features within the audio signal. Kang et al. [48] propose a jailbreak framework against MLLMs, named AdvWave. The AdvWave framework leverages a dual-phase optimization method to address the gradient shattering problem. Additionally, they also introduce an adaptive adversarial target search algorithm to dynamically construct adversarial queries targeted at specific responses. Yang et al. [120] evaluate the security of five audio-based MLLMs against various queries: (i) harmful questions in both audio and text formats, (ii) harmful questions in text format accompanied by distracting non-speech audio, and (iii) speech-specific jailbreaks. Furthermore, they also introduce a speech-specific jailbreaking method, which decomposes harmful words into letters to conceal them in the audio input, thereby bypassing the model's security protections. Xiao et al. [110] explore audio-specific edits to implement jailbreak attacks on MLLMs and introduce the Audio Editing Toolbox (AET). The AET offers a range of editing tools for audio tone adjustment, word emphasis, and noise injection. It includes edited audio datasets, which can serve as benchmark datasets for testing jailbreak attacks on MLLMs.

5.3 Multimodal Alignment

To verify the security related to modality alignment, Hu et al. [40] introduce a novel jailbreak attack algorithm, named VideoJail, which induces video-based MLLMs to amplify harmful content in images. Specifically, by leveraging carefully crafted text prompts, VideoJail can induce the model to focus its attention on malicious queries, successfully breaching the model's security mechanisms. Miao et al. [85] construct a comprehensive benchmark to evaluate the security of text-to-video models, which employ jailbreak attack-based prompts. Experimental results indicate that no single video-based MLLM excels across all critical aspects, necessitating a further trade-off between the usability and safety of the video model. Chiu et al. [17] leverage audio samples to circumvent the constraints of traditional text, breaking through the defense mechanisms of MLLMs. Specifically, they embed harmful prompts within seemingly harmless ones, reducing the MLLMs' ability to recognize and block harmful content by increasing situational complexity and introducing ambiguity. At the same time, they construct a sequence of multiple queries, placing the critical adversarial query in the middle of the sequence to reduce the risk of triggering safety mechanisms. Shen et al. [96] systematically measure the security of the GPT-4o model against jailbreak attacks. The results indicate that the GPT-4o model exhibits good resistance to malicious queries when leveraging voice mode. Furthermore, they also present a novel voice jailbreak attack algorithm that attempts to persuade the GPT-4o model through fictional storytelling. Extensive experiments validate the feasibility of this attack method.

5.4 Evaluation

Ying et al. [127] evaluate the capabilities of GPT-4o against jailbreak attacks across three modalities: text, speech, and image. Their findings expose a new attack surface for jailbreak attacks on GPT-4o specifically in the speech

modality.

Reflections and Challenges: The research discussed underscores that the inherent defense mechanisms of state-of-the-art MLLMs can be circumvented by malicious audio manipulations and queries, thereby producing harmful outputs. However, there is less research on jailbreaking attacks targeting audio-visual-based MLLMs, which requires more attention. Additionally, due to variations in model structures, jailbreaking attack algorithms may demonstrate varying performance and require substantial computational resources.

6 Other attack

Roy et al. [94] leverage the non-linear diaphragm and amplifiers of microphones to create "shadows" that facilitate covert acoustic communication. This method is applied in acoustic denial-of-service attacks. Li et al. [66] propose an algorithm for detecting audio deepfakes that operates without relying on semantic information. Specifically, the algorithm decouples semantic and acoustic information in the audio, utilizing only the acoustic information for detection. This approach effectively identifies a variety of deepfake technologies while maintaining a low error rate.

7 Defense

Compared to research on attack algorithms targeting audio-visual-based multimodal models, exploring defense algorithms appears to be more crucial [138]. Therefore, we discuss the algorithms for defending against adversarial and backdoor attacks from three perspectives: sample detection, optimization, and purification, as shown in Table 4. It is worth noting that there are no defense algorithms available for jailbreak attacks, especially for audio-visual-based multimodal models.

7.1 Adversarial attack

Sample detecting To detect adaptive adversarial audio samples, Du et al. [23] introduce a unified detection framework, which combines noise padding and sound reverberation. Specifically, an adaptive artificial speech generator is designed to enhance the effectiveness of detection. Additionally, they also design multiple noise padding strategies to disrupt the continuity of adversarial noise. The advantage of this framework is that it is applicable to various automatic speech recognition systems without the need for additional training. Kwon et al. [59] propose an adversarial audio sample detection algorithm, which adds a new low-level distortion to the samples under inspection. If the classification result is sensitive, then the sample is identified as an adversarial sample; otherwise, it is considered an original sample. Experimental results validate the effectiveness of this algorithm. Hussain et al. [44] introduce an adversarial sample detection framework named WaveGuard, which incorporates an audio transformation module. The fundamental premise of this module is that adversarial examples induce instability in model predictions, in contrast to benign examples, which maintain stability despite minor modifications to the inputs. If there's a substantial discrepancy between the transcriptions, the input is labeled as adversarial. Ma et al. [79] explore an adversarial sample detection method based on the temporal correlation for audio and video. They believe that the addition of adversarial noise can disrupt the correlation between audio and video. Therefore, a synchronization confidence score is used as a proxy to measure the correlation between audio and video, detecting adversarial samples. Kwon et al. [58] introduce a defense method for detecting adversarial examples that does not require a separate module or an additional process. Specifically, they suggest that when optimized noise is embedded into target samples, it leads to a specific pattern in the classification scores of the adversarial example. Experimental results show that this characteristic can be used to identify audio adversarial examples and defend against adversarial attacks.

To detect audio adversarial examples, Guo et al. [34] propose a universal defense algorithm based on audio fingerprint analysis. Specifically, they analyze the similarity between the current query and a set of past queries of a specified length to determine when a series of queries might be susceptible to producing audio adversarial samples. Park et al. [89] introduce an adversarial sample detection algorithm, which inputs both smoothed and original samples into the ASR system and introduces carefully selected noise to logits before decoding. According to their analysis, this carefully selected noise significantly affects the transcription results of adversarial samples but has a minimal impact on clean samples. Therefore, they leverage this characteristic to identify adversarial samples. Rabhi

Table 4 Summary of defense algorithms against audio-visual attacks, which includes the method name, capability, defense characteristics, representative models and tasks, and partial contributions.

Method	Capability	Characteristics	Model	Task	Contribution
WaveGuard [23]	White/Gray-box	Sample detecting	Deepspeech	Speech Recognition	Detects audio adversarial examples without model retraining
CS-Detect [59]	White/Gray-box	Sample detecting	RNN	Speech Recognition	Detects audio adversarial examples using classification score patterns
Agitated [89]	White/Black-box	Sample detecting	RNN	Speech Recognition	Introduces noise to logits for adversarial example detection
Robust [25]	White/Black-box	Optimization	SVM	Audio Classification	Uses DWT and SVM for robust audio attack defense
SNR [76]	White-box	Optimization	ResNet18	Audio Classification	Applies SNR, APGD attacks, and Cutmix for robustness improvement
HF-Smoothing [87]	White-box	Purification	CNN	Speaker Identification	Utilizes high-frequency smoothing to enhance adversarial robustness
FeCo [10]	White/Black-box	Purification	CNN	Speaker Recognition	Combines feature compression with adversarial training for robust defenses
AudioPure [107]	White/Black-box	Purification	Diffusion	Speech Command Recognition	Uses diffusion models for effective adversarial audio purification
SpeechGuard [114]	White/Black-box	Sample detecting	Autoencoder	Speech Recognition	Combines detection and purification for robust backdoor attack defense
Sniper [35]	Black-box	Sample detecting	t-SNE	Speaker Verification	Effectively cleanses datasets using sniper-based defense mechanism
GN-FT [148]	White-box	Optimization	LSTM	Speech Recognition	Reduces backdoored effects by penalizing high-gradient neurons
Silero-VAD [3]	White-box	Purification	Transformer	Speech Recognition	Uses VAD to filter malicious triggers and protect models
KDDF [15]	White-box	Purification	ResNeXt	Speech Recognition	Detects triggers and recovers poisoned audio data using KD.

et al. [93] demonstrate that the audio deepfake detection system is vulnerable to adversarial examples. To counter these threats, they introduce a highly generalizable defense mechanism that includes a speech-to-text mechanism. Specifically, if the audio is classified as real by the deepfake detector, the speech-to-text function evaluates whether the audio content matches the expected text. If the audio successfully passes this speech-to-text verification layer, it is then deemed authentic.

Optimization Esmaeilpour et al. [25] introduce a defense algorithm against adversarial attacks. They first demonstrate the robustness of support vector machines (SVMs) when facing several state-of-the-art adversarial attacks. Then, they design a new method based on audio signal preprocessing with Discrete Wavelet Transform (DWT) representations and SVM to protect audio systems from adversarial attacks, which provides a viable balance between accuracy and robustness. Lu et al. [76] enhance the adversarial robustness of audio classifiers through three aspects. First, they demonstrate that ℓ_2 norm perturbations can generate perturbed examples with specific signal-to-noise ratios. Second, the APGD method is introduced for adversarial training, which enhances the model's robustness against adversarial attacks. Lastly, they leverage data augmentation strategies, such as CutMix, to optimize the model's robustness. Sun et al. [98] leverage MLLMs to build an audio-guided navigation agent, which evaluates the credibility of human instructions based on emotional cues in spoken communication, such as tone and intonation variations. Experimental results indicate that the system possesses remarkable resilience when facing adversarial attacks.

Purification B. Raj et al. [87] present a more robust adversarial attack defense algorithm than naive noise filtering. Specifically, they utilize a high-pass filter on additive Gaussian noise to smooth the model where it is most vulnerable. Chen et al. [10] provide valuable insights for enhancing the security of speaker recognition systems through a comprehensive evaluation of various transformation and training methods. They also introduce a new feature compression technique, named FeCo, which compresses and aggregates audio features using clustering methods to mitigate adversarial perturbations, ensuring the security of speaker recognition systems. Wu et al. [107] introduce an adversarial purification-based defense algorithm based on off-the-shelf diffusion models. This algorithm adds a small amount of noise to the adversarial audio, then runs the reverse sampling step to purify the noisy audio and recover clean audio. It is plug-and-play and can be quickly transferred to other models without the need for retraining. Du et al. [24] introduce an adaptive unified defense framework tailored to adversarial attacks. Specifically, they design a unified pre-processing mechanism, which includes RIR convolution, multi-fragment noise padding, and SPL complexity analysis, to disrupt adversarial attacks. Secondly, they leverage multiple automatic speech recognition systems to transcribe pre-processed audio to evaluate similarity and detect adversarial properties. Experimental results show that the framework effectively defends against various adversarial attack strategies.

Insights: The research discussed highlights various adversarial sample detection and purification strategies, such as noise padding and audio fingerprint analysis, which enhance the adaptability and robustness of audio-visual models against different types of attacks. Despite the multitude of defense algorithms, these methods lack sufficient generalizability when facing various types of attacks. Additionally, defending against adversarial attacks requires substantial computational resources, which reduces their practicality.

7.2 Backdoor attack

Sample detecting Building upon STRIP [29], a backdoor attack defense algorithm targeting images, Xin et al. [114] introduce SpeechGuard to defend against backdoor attacks in the audio domain. Specifically, they optimize the STRIP algorithm by enhancing the perturbation techniques, making it more effective for detecting poisoned audio samples. Furthermore, they utilize time-frequency masking to suppress trigger signals and purify the poisoned samples. Guo et al. [35] introduce a "sniper"-based backdoor attack defense algorithm that examines the dataset before training to filter out suspicious samples. Specifically, they use the average embedding of the dataset as a "sniper" and calculate the ℓ_2 distance between it and other samples to identify potentially malicious samples.
Optimization Zhou et al. [148] discover that backdoored neurons exhibit greater gradient values compared to other neurons. Based on this observation, they propose a gradient norm-based fine-tuning algorithm to defend against backdoor attacks. Specifically, they apply gradient norm regularization to fine-tune the victim model, thereby weakening the backdoored neurons and enhancing the model's defense against backdoor attacks.
Purification Zhu et al. [149] introduce a backdoor attack defense algorithm for voice print recognition models based on speech enhancement and weight pruning. Specifically, this algorithm leverages perturbation detection to distinguish between clean and poisoned samples. Subsequently, clean samples are used to fine-tune the model with pruning, and a speech enhancement algorithm is applied to purify the poisoned samples, thereby preventing the activation of the backdoor. Chen et al. [15] present a backdoor attack defense algorithm in federated learning. Specifically, they detect and remove features of the trigger during inference based on knowledge distillation. The knowledge distillation algorithm is used to train a validation model on each IoT device to identify suspicious poisoned samples, and a feature cancellation mechanism is employed to eliminate the trigger features in these suspicious samples, thereby defending against backdoor attacks. Wu et al. [109] design a backdoor attack defense algorithm for federated learning systems, which is based on speaker frequency. This algorithm first filters out infrasound frequencies below 18 Hz and ultrasonic frequencies above 20,000 Hz, aiming to eliminate commonly used backdoor attack triggers. Then, by filtering out frequencies outside the normal human vocal range, they ensure that the integrity and quality of the voice data are preserved, and effectively prevent the activation of backdoors. Bartolini et al. [3] explore leveraging the voice activity detection model as a defense mechanism against backdoor attacks, aiming to filter out backdoor triggers and prevent the backdoor from being activated. Experimental results indicate that this model is capable of defending against backdoor attacks.

> **Issues to Consider:** Due to the fixed trigger patterns inherent in backdoor attacks, these patterns can serve as distinctive signals for identification and subsequent defense mechanisms. However, the challenge in defending against such attacks extends beyond merely detecting poisoned audio samples. It is equally crucial to ensure that clean audio-visual samples remain unaffected. This dual requirement significantly amplifies the complexity of implementing effective defense strategies against backdoor attacks.

7.3 Other

Yu et al. [129] introduce AntiFake, a defense strategy to prevent unauthorized speech synthesis. Concretely, AntiFake disrupts speech synthesis by deviating speaker embeddings used for speaker identity control. Furthermore, to ensure transferability, they also leverage ensemble learning to enhance the generalizability of the AntiFake algorithm.

8 Application

Much like a coin has two sides, various attacks on audio-visual-based multimodal models can serve both as a threat and as an effective tool for evaluation. This dual role facilitates activities such as model copyright auditing and privacy protection, highlighting the complexity and multifaceted nature of security in audio-visual-based multimodal models.
Copyright protection To protect the copyright of speech recognition models, Liao et al. [69] introduce imperceptible backdoor watermarks to authenticate ownership. They utilize Gaussian noise watermarks, extreme frequency Gaussian noise watermarks, and unrelated audio watermarks to embed backdoors into the target model. This enables black-box verification of the intellectual property of the model owners. Wu et al. [106] propose a steganography method based on adversarial examples, which involves making subtle perturbations to audio in the time domain to evade detection by CNN steganalysis. This method does not rely on existing embedding costs but instead starts with

random or simple embeddings, enhancing security performance through iterative cost updates. Chen et al. [11] introduce a model access control scheme based on hidden adversarial samples, focused on the intellectual property protection of automatic speech recognition models. This method utilizes audio adversarial samples by embedding user identity information into the adversarial samples of the audio, serving as proof samples for authentication.

Privacy To protect user privacy, O'Reilly et al. [88] propose VoiceBlock, a system that performs adversarial modifications on user audio streams in real-time, de-identifying user speech to safeguard privacy from automatic speaker recognition systems. VoiceBlock employs deep networks and applies time-varying finite impulse response filters to the outgoing audio stream. These modifications prevent automated systems from identifying users based on their voice, while retaining the speech characteristics unchanged for human listeners.

Dataset To better align fake speech datasets with real-life scenarios, Luong et al. [78] hconstruct LlamaPartialSpoof, a 130-hour dataset that contains both fully and partially fake speech. This dataset is created using MLLMs and voice cloning technologies. Its purpose is to provide a more comprehensive assessment of the vulnerabilities in current countermeasure systems. Cai et al. [7] construct a dataset named AV-Deepfake1M. This dataset includes three types of manipulations: video manipulations, audio manipulations, and audio-visual manipulations. It addresses the gap in existing deepfake datasets, which do not include these types of partial manipulations.

9 Summary and Challenges

In this section, we summarize the similarities and differences between various types of attacks and defense algorithms, with the aim of gaining a deeper understanding of model security. Additionally, we provide a summary of the challenges and trends to further encourage research into model security.

9.1 Summary of attacks and defenses

Similarity Analysis Despite the diverse forms of attack algorithms, their goal is to maliciously manipulate the responses of audio-visual-based multimodal models. Especially, adversarial and jailbreak attacks can influence the outputs of models through multiple iterative queries. Therefore, different attack methods exhibit elements that are instructive for reciprocal learning. For example, perturbations in adversarial attacks have the potential to facilitate jailbreak attacks, thereby disrupting the internal defense mechanisms of models and resulting in the output of harmful content. Moreover, the defense algorithms against adversarial and backdoor attacks can be categorized into sample detection, optimization, and purification, which can also serve as defensive strategies against jailbreak attacks, particularly through purification.

Difference Analysis Unlike adversarial or jailbreak attacks, most backdoor attack algorithms require fine-tuning the model to establish an association between the trigger and the target label, thus, backdoor attacks may consume more computational resources in MLLMs. Furthermore, alterations in the fine-tuned model weights may detrimentally affect the model's generalization capabilities. Consequently, attackers must meticulously balance the effectiveness of backdoor attacks with the preservation of the model's standard performance. In contrast, the primary focus of adversarial attacks is on refining the generation of perturbations to reduce the frequency of queries.

9.2 Challenges and trends

Attack Algorithms Despite the proliferation of audio-visual attack algorithms, several challenges remain:

- Facing different MLLM architectures, the generalizability of adversarial perturbations significantly impacts the attack success rate. Although previous research has continuously optimized the generalizability of perturbations, the evolving defense capabilities of MLLMs, which leverage security-aligned optimizations, mean that constructing more effective and generalized perturbations remains an ongoing challenge.
- For adversarial and jailbreak attacks, particularly in real-time audio streaming systems, attackers are required to make multiple queries to achieve their ultimate objectives. However, in closed-source models, repeated queries require multiple API calls, which can lead to significant expenses. Therefore, exploring efficient algorithms for generating malicious queries or adversarial perturbations is a worthwhile endeavor.
- As the parameter count of models increases, backdoor attack algorithms that depend on fine-tuning are becoming impractical. Therefore, exploring innovative backdoor attack algorithms that either do not require fine-tuning or only utilize parameter-efficient fine-tuning, such as poisoning-based instruction or in-context learning, is encouraged. For instance, Zhao et al. [140] propose ICLAttack method based on in-context learning algorithms, avoiding the fine-tuning of large language models. This method can be adapted for use in audio-visual attacks.

• In existing backdoor attack algorithms, to establish an alignment between the trigger and the target label, attackers need to implant triggers and modify sample labels. Although this can achieve a feasible attack success rate, samples with modified labels may be detected by defense algorithms, which reduces the stealthiness of the backdoor attack. Therefore, exploring backdoor attack algorithms that do not require modifying sample labels is a trend.

• Moreover, exploring the interpretability of attack algorithms is crucial for understanding the deeper working mechanisms of attacks and can also enhance our understanding of the security vulnerabilities in MLLMs. Specifically, interpretable jailbreak attack algorithms can play a crucial role in identifying and addressing the inherent defensive weaknesses of models.

Defense Algorithms

• The purpose of exploring attack algorithms is to identify potential security vulnerabilities in audio-visual-based multimodal models, thereby facilitating the development of effective defense algorithms. However, existing defense algorithms are limited to specific types of attacks, which lack sufficient generalization performance, reducing their effectiveness. Therefore, the exploration of more effective defense algorithms remains an ongoing challenge.

• In real-time audio streaming systems, which operate under strict time constraints, this challenge becomes particularly pronounced. In MLLM-based systems, the large number of parameters often results in the majority of the time budget being consumed by model inference, leaving only a limited window for rejecting or filtering malicious inputs. This restriction may compromise the effectiveness of defense algorithms. Striking a balance between model efficiency and security thus emerges as a critical challenge for real-time audio streaming systems.

• Effective defense algorithms are essential not only for the accurate identification of poisoned audio samples but also for minimizing false positive rates, which are crucial for maintaining the normal performance of models. Particularly in the context of sample detection algorithms, the ability of defense mechanisms to precisely differentiate between poisoned and original audio samples is fundamental to the overall reliability and stability of the model.

• Although optimization-based algorithms demonstrate notable efficacy in defending against attacks, they require the fine-tuning of the victim model, which in turn demands substantial computational resources, particularly within the framework of MLLMs. Therefore, reducing the consumption of computational resources during the defense phase and enhancing the efficiency of backdoor mitigation are imperative for evaluating the practicality of defense algorithms.

• The absence of defense algorithms against jailbreak attacks targeting audio-visual-based MLLMs presents an urgent challenge. A feasible approach is to leverage defense algorithms from the image or text domains. For example, Wang et al. define the defense against jailbreak attacks as "backdoor attacks", where prefixed safety examples with a secret prompt serve as "triggers". By fine-tuning, they establish an alignment between the triggers and safety responses, thus preventing the output of harmful content to defend against jailbreak attacks. The aforementioned algorithm can be adapted to defend against jailbreak attacks targeting audio-based tasks.

10 Conclusion

In this paper, we systematically review various attack and defense algorithms in the audio-visual domain. Simultaneously, we focus our attention on multimodal large language models (MLLMs), which reveal the potential security vulnerabilities of MLLMs. Our research reveals that state-of-the-art audio-visual-based multimodal models are susceptible to adversarial perturbations or malicious queries, resulting in the output of incorrect or harmful content. In addition, we demonstrate existing defense algorithms against attacks, with sample detection, optimization, and purification. Finally, we highlight the potential challenges and emerging trends in audio-visual attacks. We hope that this survey serves as a valuable resource for researchers and practitioners, fostering the security of MLLMs and building a reliable audio-visual community.

References

1 Abdoli, S., Hafemann, L.G., Rony, J., Ayed, I.B., Cardinal, P., Koerich, A.L.: Universal adversarial audio perturbations. arXiv preprint arXiv:1908.03173 (2019)

2 Al Kader Hammoud, H.A., Liu, S., Alkhrashi, M., Albalawi, F., Ghanem, B.: Look listen and attack: Backdoor attacks against video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3439–3450 (2024)

3 Bartolini, J., Stoyanov, T., Giaretta, A.: Hidden in plain sound: Environmental backdoor poisoning attacks on whisper, and mitigations. arXiv preprint arXiv:2409.12553 (2024)

4 Cai, H., Zhang, P., Dong, H., Xiao, Y., Ji, S.: Pbsm: backdoor attack against keyword spotting based on pitch boosting and sound masking. arXiv preprint arXiv:2211.08697 (2022)

5 Cai, H., Zhang, P., Dong, H., Xiao, Y., Ji, S.: Vsvc: backdoor attack against keyword spotting based on voiceprint selection and voice conversion. arXiv preprint arXiv:2212.10103 (2022)

6 Cai, H., Zhang, P., Dong, H., Xiao, Y., Koffas, S., Li, Y.: Towards stealthy backdoor attacks against speech recognition via elements of sound. IEEE Transactions on Information Forensics and Security (2024)
7 Cai, Z., Ghosh, S., Adatia, A.P., Hayat, M., Dhall, A., Gedeon, T., Stefanov, K.: Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7414–7423 (2024)
8 Carlini, N., Wagner, D.: Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW). pp. 1–7. IEEE (2018)
9 Chang, K.H., Huang, P.H., Yu, H., Jin, Y., Wang, T.C.: Audio adversarial examples generation with recurrent neural networks. In: 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC). pp. 488–493. IEEE (2020)
10 Chen, G., Zhao, Z., Song, F., Chen, S., Fan, L., Wang, F., Wang, J.: Towards understanding and mitigating audio adversarial examples for speaker recognition. IEEE Transactions on Dependable and Secure Computing **20**(5), 3970–3987 (2022)
11 Chen, H., Zhang, J., Chen, K., Zhang, W., Yu, N.: Model access control based on hidden adversarial examples for automatic speech recognition. IEEE Transactions on Artificial Intelligence **5**(3), 1302–1315 (2023)
12 Chen, M., Lu, L., Yu, J., Ba, Z., Lin, F., Ren, K.: Advreverb: Rethinking the stealthiness of audio adversarial examples to human perception. IEEE Transactions on Information Forensics and Security **19**, 1948–1962 (2023)
13 Chen, M., Xu, X., Lu, L., Ba, Z., Lin, F., Ren, K.: Devil in the room: triggering audio backdoors in the physical world. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 7285–7302 (2024)
14 Chen, Q., Chen, M., Lu, L., Yu, J., Chen, Y., Wang, Z., Ba, Z., Lin, F., Ren, K.: Push the limit of adversarial example attack on speaker recognition in physical domain. In: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. pp. 710–724 (2022)
15 Chen, Y.W., Ke, B.H., Chen, B.Z., Chiu, S.R., Tu, C.W., Kuo, J.J.: Knowledge distillation based defense for audio trigger backdoor in federated learning. In: GLOBECOM 2023-2023 IEEE Global Communications Conference. pp. 4271–4276. IEEE (2023)
16 Cheng, P., Wang, Y., Huang, P., Ba, Z., Lin, X., Lin, F., Lu, L., Ren, K.: Alif: Low-cost adversarial audio attacks on black-box speech platforms using linguistic features. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). pp. 1628–1645. IEEE (2024)
17 Chiu, C.W., Huang, L., Li, B., Chen, H.: Do as i say not as i do': A semi-automated approach for jailbreak prompt attack against multimodal llms. arXiv e-prints pp. arXiv–2502 (2025)
18 Choi, H., Jung, J.H., Yoon, J.W.: Ghost in the radio: An audio adversarial attack using environmental noise through radio. IEEE Access (2024)
19 Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M., Zhang, Y.: Comprehensive assessment of jailbreak attacks against llms. arXiv preprint arXiv:2402.05668 (2024)
20 Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., Zhou, J.: Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919 (2023)
21 Dou, Z., Hu, X., Yang, H., Liu, Z., Fang, M.: Adversarial attacks to multi-modal models. In: Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis. pp. 35–46 (2023)
22 Du, X., Pun, C.M.: Robust audio patch attacks using physical sample simulation and adversarial patch noise generation. IEEE Transactions on Multimedia **24**, 4381–4393 (2021)
23 Du, X., Pun, C.M., Zhang, Z.: A unified framework for detecting audio adversarial examples. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3986–3994 (2020)
24 Du, X., Zhang, Q., Zhu, J., Liu, X.: Adaptive unified defense framework for tackling adversarial audio attacks. Artificial Intelligence Review **57**(8), 218 (2024)
25 Esmaeilpour, M., Cardinal, P., Koerich, A.L.: A robust approach for securing audio classification against adversarial attacks. IEEE Transactions on Information Forensics and Security **15**, 2147–2159 (2019)
26 Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y.: Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666 (2024)
27 Fang, Z., Wang, T., Zhao, L., Zhang, S., Li, B., Ge, Y., Li, Q., Shen, C., Wang, Q.: Zero-query adversarial attack on black-box automatic speech recognition systems. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. pp. 630–644 (2024)
28 Farooq, M.U., Khan, A., Uddin, K., Malik, K.M.: Transferable adversarial attacks on audio deepfake detection. arXiv preprint arXiv:2501.11902 (2025)
29 Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S.: Strip: A defence against trojan attacks on deep neural networks. In: Proceedings of the 35th annual computer security applications conference. pp. 113–125 (2019)
30 Ge, Y., Zhao, L., Wang, Q., Duan, Y., Du, M.: Advddos: Zero-query adversarial attacks against commercial speech recognition systems. IEEE Transactions on Information Forensics and Security **18**, 3647–3661 (2023)
31 Gong, C., Wang, X., Cooper, E., Wells, D., Wang, L., Dang, J., Richmond, K., Yamagishi, J.: Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
32 Gong, T., Ramos, A.G.C., Bhattacharya, S., Mathur, A., Kawsar, F.: Audidos: Real-time denial-of-service adversarial attacks on deep audio models. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). pp. 978–985. IEEE (2019)
33 Gong, X., Fang, Z., Li, B., Wang, T., Chen, Y., Wang, Q.: Palette: Physically-realizable backdoor attacks against video recognition models. IEEE Transactions on Dependable and Secure Computing **21**(4), 2672–2685 (2023)
34 Guo, F., Sun, Z., Chen, Y., Ju, L.: Towards the universal defense for query-based audio adversarial attacks on speech recognition system. Cybersecurity **6**(1), 40 (2023)
35 Guo, H., Chen, X., Guo, J., Xiao, L., Yan, Q.: Masterkey: Practical backdoor attack against speaker verification systems. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. pp. 1–15 (2023)
36 Guo, H., Wang, G., Wang, Y., Chen, B., Yan, Q., Xiao, L.: Phantomsound: Black-box, query-efficient audio adversarial attack via split-second phoneme injection. In: Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses. pp. 366–380 (2023)
37 Guo, H., Wang, Y., Ivanov, N., Xiao, L., Yan, Q.: Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. In: Proceedings of the 2022 ACM SIGSAC conference on computer and communications security. pp. 1353–1366 (2022)
38 Gupta, I., Khachaturov, D., Mullins, R.: " i am bad": Interpreting stealthy, universal and robust audio jailbreaks in audio-language models. arXiv preprint arXiv:2502.00718 (2025)
39 Han, X., Wu, Y., Zhang, Q., Zhou, Y., Xu, Y., Qiu, H., Xu, G., Zhang, T.: Backdooring multimodal learning. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 3385–3403. IEEE (2024)
40 Hu, W., Gu, S., Wang, Y., Hong, R.: Videojail: Exploiting video-modality vulnerabilities for jailbreak attacks on multimodal large language models. In: ICLR 2025 Workshop on Building Trust in Language Models and Applications (2025)
41 Huang, L., Jiang, X., Wang, Z., Mo, W., Xiao, X., Han, B., Yin, Y., Zheng, F.: Image-based multimodal models as intruders: Transferable multimodal attacks on video-based mllms. arXiv preprint arXiv:2501.01042 (2025)
42 Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., Sharma, M.: Best-of-n jailbreaking. arXiv preprint arXiv:2412.03556 (2024)
43 Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

44 Hussain, S., Neekhara, P., Dubnov, S., McAuley, J., Koushanfar, F.: Waveguard: Understanding and mitigating audio adversarial examples. In: 30th USENIX security symposium (USENIX Security 21). pp. 2273–2290 (2021)
45 Jia, Y., Wu, X., Li, H., Zhang, Q., Hu, Y., Zhao, S., Fan, W.: Uni-retrieval: A multi-style retrieval framework for stem's education. arXiv preprint arXiv:2502.05863 (2025)
46 Jia, Y., Wu, X., Zhang, Q., Qin, Y., Xiao, L., Zhao, S.: Towards robust evaluation of stem education: Leveraging mllms in project-based learning. arXiv preprint arXiv:2505.17050 (2025)
47 Jia, Y., Xie, J., Jivaganesh, S., Li, H., Wu, X., Zhang, M.: Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. arXiv preprint arXiv:2505.11217 (2025)
48 Kang, M., Xu, C., Li, B.: Advwave: Stealthy adversarial jailbreak attack against large audio-language models. In: The Thirteenth International Conference on Learning Representations (2024)
49 Kasher, M., Zhao, M., Greenberg, A., Gulati, D., Kokalj-Filipovic, S., Spasojevic, P.: Inaudible manipulation of voice-enabled devices through backdoor using robust adversarial audio attacks. In: Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. pp. 37–42 (2021)
50 Keller, L., Glavaš, G.: Speechtaxi: On multilingual semantic speech classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025)
51 Kheddar, H., Hemis, M., Himeur, Y.: Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion p. 102422 (2024)
52 Kim, H., Park, J., Lee, J.: Generating transferable adversarial examples for speech classification. Pattern Recognition **137**, 109286 (2023)
53 Ko, K., Kim, S., Kwon, H.: Multi-targeted audio adversarial example for use against speech recognition systems. Computers & Security **128**, 103168 (2023)
54 Koffas, S., Pajola, L., Picek, S., Conti, M.: Going in style: Audio backdoors through stylistic transformations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)
55 Koffas, S., Xu, J., Conti, M., Picek, S.: Can you hear it? backdoor attacks via ultrasonic triggers. In: Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning. pp. 57–62 (2022)
56 Kong, Y., Zhang, J.: Adversarial audio: A new information hiding method. In: Proc. Interspeech 2020. pp. 2287–2291 (2020)
57 Kwon, H., Kim, Y., Yoon, H., Choi, D.: Selective audio adversarial example in evasion attack on speech recognition system. IEEE Transactions on Information Forensics and Security **15**, 526–538 (2019)
58 Kwon, H., Nam, S.H.: Audio adversarial detection through classification score on speech recognition systems. Computers & Security **126**, 103061 (2023)
59 Kwon, H., Yoon, H., Park, K.W.: Poster: Detecting audio adversarial example through audio modification. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. pp. 2521–2523 (2019)
60 Lan, J., Wang, J., Yan, B., Yan, Z., Bertino, E.: Flowmur: A stealthy and practical audio backdoor attack with limited knowledge. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). pp. 1646–1664. IEEE (2024)
61 Lan, J., Zhang, R., Yan, Z., Wang, J., Chen, Y., Hou, R.: Adversarial attacks and defenses in speaker recognition systems: A survey. Journal of Systems Architecture **127**, 102526 (2022)
62 Lee, Y., Chen, K., Meng, G., Lv, P., et al.: Aliasing backdoor attacks on pre-trained models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2707–2724 (2023)
63 Li, H., Jia, P., Li, W., Ma, B., Li, B., Wu, D., Li, H.: Towards efficient universal adversarial attack on audio classification models: A two-step method. In: International Symposium on Emerging Information Security and Applications. pp. 20–37. Springer (2023)
64 Li, J., Gao, K., Bai, Y., Zhang, J., Xia, S.t., Wang, Y.: Fmm-attack: A flow-based multi-modal adversarial attack on video-based llms. arXiv preprint arXiv:2403.13507 (2024)
65 Li, J., Qu, S., Li, X., Szurley, J., Kolter, J.Z., Metze, F.: Adversarial music: Real world audio adversary against wake-word detection system. Advances in Neural Information Processing Systems **32** (2019)
66 Li, X., Li, K., Zheng, Y., Yan, C., Ji, X., Xu, W.: Safeear: Content privacy-preserving audio deepfake detection. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. pp. 3585–3599 (2024)
67 Li, X., Ze, J., Yan, C., Cheng, Y., Ji, X., Xu, W.: Enrollment-stage backdoor attacks on speaker recognition systems via adversarial ultrasound. IEEE Internet of Things Journal **11**(8), 13108–13124 (2023)
68 Li, Z., Shi, C., Xie, Y., Liu, J., Yuan, B., Chen, Y.: Practical adversarial attacks against speaker recognition systems. In: Proceedings of the 21st international workshop on mobile computing systems and applications. pp. 9–14 (2020)
69 Liao, J., Yi, L., Shi, W., Yang, W., Fang, Y., Yang, X.: Imperceptible backdoor watermarks for speech recognition model copyright protection. Visual Intelligence **2**(1), 23 (2024)
70 Liu, M., Li, X., Wang, M., Zhang, X.L., Rahardja, S.: Mtbv: Multi-trigger backdoor attacks on speaker verification. In: 2024 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). pp. 1–5. IEEE (2024)
71 Liu, P., Zhang, S., Yao, C., Ye, W., Li, X.: Backdoor attacks against deep neural networks by personalized audio steganography. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 68–74. IEEE (2022)
72 Liu, Q., Zhou, T., Cai, Z., Tang, Y.: Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2390–2398 (2022)
73 Liu, T., Lin, F., Wang, Z., Wang, C., Ba, Z., Lu, L., Xu, W., Ren, K.: Magbackdoor: Beware of your loudspeaker as a backdoor for magnetic injection attacks. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 3416–3431. IEEE (2023)
74 Liu, X., Wan, K., Ding, Y., Zhang, X., Zhu, Q.: Weighted-sampling audio adversarial example attack. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4908–4915 (2020)
75 Liu, X., Cui, X., Li, P., Li, Z., Huang, H., Xia, S., Zhang, M., Zou, Y., He, R.: Jailbreak attacks and defenses against multimodal generative models: A survey. arXiv preprint arXiv:2411.09259 (2024)
76 Lu, K., Nguyen, M.C., Xu, X., Foo, C.S.: On adversarial robustness of audio classifiers. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)
77 Luo, Y., Tai, J., Jia, X., Zhang, S.: Practical backdoor attack against speaker recognition system. In: International Conference on Information Security Practice and Experience. pp. 468–484. Springer (2022)
78 Luong, H.T., Li, H., Zhang, L., Lee, K.A., Chng, E.S.: Llamapartialspoof: An llm-driven fake speech dataset simulating disinformation generation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025)
79 Ma, P., Petridis, S., Pantic, M.: Detecting adversarial attacks on audiovisual speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6403–6407 (2021)
80 Mao, J., Zhu, S., Liu, J.: An inaudible voice attack to context-based device authentication in smart iot systems. Journal of Systems Architecture **104**, 101696 (2020)
81 Mao, Y., Liu, P., Cui, T., Liu, C., You, D.: Divide and conquer: A hybrid strategy defeats multimodal large language models. arXiv preprint arXiv:2412.16555 (2024)
82 Mengara, O.: The art of deception: Robust backdoor attack using dynamic stacking of triggers. arXiv preprint arXiv:2401.01537 (2024)
83 Mengara, O.: A backdoor approach with inverted labels using dirty label-flipping attacks. IEEE Access (2024)

84 Mengara, O.: Trading devil: Robust backdoor attack via stochastic investment models and bayesian approach. arXiv preprint arXiv:2406.10719 (2024)
85 Miao, Y., Zhu, Y., Yu, L., Zhu, J., Gao, X.S., Dong, Y.: T2vsafetybench: Evaluating the safety of text-to-video generative models. Advances in Neural Information Processing Systems **37**, 63858–63872 (2024)
86 Mun, H., Seo, S., Son, B., Yun, J.: Black-box audio adversarial attack using particle swarm optimization. IEEE Access **10**, 23532–23544 (2022)
87 Olivier, R., Raj, B., Shah, M.: High-frequency adversarial defense for speech and audio. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2995–2999 (2021)
88 O'Reilly, P., Bugler, A., Bhandari, K., Morrison, M., Pardo, B.: Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. Advances in Neural Information Processing Systems **35**, 30058–30070 (2022)
89 Park, N., Kim, J.: Toward robust asr system against audio adversarial examples using agitated logit. ACM Transactions on Privacy and Security **27**(2), 1–26 (2024)
90 Qi, G., Chen, Y., Zhu, Y., Hui, B., Li, X., Mao, X., Zhang, R., Xue, H.: Transaudio: Towards the transferable adversarial audio attack via learning contextualized perturbations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023)
91 Qiu, S., You, X., Rong, W., Huang, L., Liang, Y.: Boosting imperceptibility of adversarial attacks for environmental sound classification. In: 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 790–797. IEEE (2024)
92 Qu, X., Wei, P., Gao, M., Sun, Z., Ong, Y.S., Ma, Z.: Synthesising audio adversarial examples for automatic speech recognition. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1430–1440 (2022)
93 Rabhi, M., Bakiras, S., Di Pietro, R.: Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications **250**, 123941 (2024)
94 Roy, N., Hassanieh, H., Roy Choudhury, R.: Backdoor: Making microphones hear inaudible sounds. In: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. pp. 2–14 (2017)
95 Schoof, C., Koffas, S., Conti, M., Picek, S.: Emoback: Backdoor attacks against speaker identification using emotional prosody. In: Proceedings of the 2024 Workshop on Artificial Intelligence and Security. pp. 137–148 (2024)
96 Shen, X., Wu, Y., Backes, M., Zhang, Y.: Voice jailbreak attacks against gpt-4o. arXiv preprint arXiv:2405.19103 (2024)
97 Shi, C., Zhang, T., Li, Z., Phan, H., Zhao, T., Wang, Y., Liu, J., Yuan, B., Chen, Y.: Audio-domain position-independent backdoor attack via unnoticeable triggers. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. pp. 583–595 (2022)
98 Sun, X., Zhang, Y., Tang, X., Bedi, A.S., Bera, A.: Trustnavgpt: Modeling uncertainty to improve trustworthiness of audio-guided llm-based robot navigation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8794–8801. IEEE (2024)
99 Takahashi, N., Inoue, S., Mitsufuji, Y.: Adversarial attacks on audio source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 521–525 (2021)
100 Tang, Y., Sun, L., Xu, X.: Silenttrig: An imperceptible backdoor attack against speaker identification with hidden triggers. Pattern Recognition Letters **177**, 103–109 (2024)
101 Taori, R., Kamsetty, A., Chu, B., Vemuri, N.: Targeted adversarial examples for black box audio systems. In: Proceedings of the 2019 IEEE security and privacy workshops (SPW). pp. 15–20. IEEE (2019)
102 Tong, C., Zheng, X., Li, J., Ma, X., Gao, L., Xiang, Y.: Query-efficient black-box adversarial attacks on automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3981–3992 (2023)
103 Vadillo, J., Santana, R.: On the human evaluation of universal audio adversarial perturbations. Computers & Security **112**, 102495 (2022)
104 Wang, J., Chen, Z., Yin, Z., Yang, Q., Liu, X.: Phonemic adversarial attack against audio recognition in real world. arXiv preprint arXiv:2211.10661 (2022)
105 Wang, S., Zhang, Z., Zhu, G., Zhang, X., Zhou, Y., Huang, J.: Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems. IEEE Transactions on Information Forensics and Security **18**, 351–364 (2022)
106 Wu, J., Chen, B., Luo, W., Fang, Y.: Audio steganography based on iterative adversarial attacks against convolutional neural networks. IEEE transactions on information forensics and security **15**, 2282–2294 (2020)
107 Wu, S., Wang, J., Ping, W., Nie, W., Xiao, C.: Defending against adversarial audio via diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
108 Wu, X., Ma, S., Shen, C., Lin, C., Wang, Q., Li, Q., Rao, Y.: Kenku: Towards efficient and stealthy black-box adversarial attacks against asr systems. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 247–264 (2023)
109 Wu, Y., Chen, J., Lei, T., Yu, J., Hossain, M.S.: Web 3.0 security: Backdoor attacks in federated learning-based automatic speaker verification systems in the 6g era. Future Generation Computer Systems **160**, 433–441 (2024)
110 Xiao, E., Cheng, H., Shao, J., Duan, J., Xu, K., Yang, L., Gu, J., Xu, R.: Tune in, act up: Exploring the impact of audio modality-specific edits on large audio language models in jailbreak. arXiv preprint arXiv:2501.13772 (2025)
111 Xiao, L., Mao, R., Zhao, S., Lin, Q., Jia, Y., He, L., Cambria, E.: Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis. IEEE Transactions on Affective Computing (2025)
112 Xiao, Y., Yao, W., Li, Z., Yang, J., Wen, W.: Phoneme semantic backdoor attacks with multiple task learning for speech classification task. In: National Conference on Man-Machine Speech Communication. pp. 79–90. Springer (2024)
113 Xie, Y., Li, Z., Shi, C., Liu, J., Chen, Y., Yuan, B.: Enabling fast and universal audio adversarial attack using generative model. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 35, pp. 14129–14137 (2021)
114 Xin, J., Lv, X.: Speechguard: Online defense against backdoor attacks on speech recognition models. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
115 Xin, J., Lyu, X., Ma, J.: Natural backdoor attacks on speech recognition models. In: International Conference on Machine Learning for Cyber Security. pp. 597–610. Springer (2022)
116 Xiong, B., Xing, Z., Wen, W.: Phoneme substitution: A novel approach for backdoor attacks on speech recognition systems. In: 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 540–547. IEEE (2024)
117 Xu, W., Xu, Y., Zhang, S.: Sample-independent federated learning backdoor attack in speaker recognition. Cluster Computing **28**(3), 158 (2025)
118 Xu, Z., Liu, Y., Deng, G., Li, Y., Picek, S.: A comprehensive study of jailbreak attack versus defense for large language models. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 7432–7449 (2024)
119 Yan, B., Lan, J., Yan, Z.: Backdoor attacks against voice recognition systems: A survey. ACM Computing Surveys **57**(3), 1–35 (2024)
120 Yang, H., Qu, L., Shareghi, E., Haffari, G.: Audio is the achilles' heel: Red teaming audio large multimodal models. arXiv preprint arXiv:2410.23861 (2024)
121 Yang, W., Li, Y., Fang, M., Wei, Y., Zhou, T., Chen, L.: Who can withstand chat-audio attacks? an evaluation benchmark for large language models. arXiv preprint arXiv:2411.14842 (2024)
122 Yao, W., Chen, Z.X., Liu, J., Wen, W., et al.: Emoattack: Utilizing emotional voice conversion for speech backdoor attacks on deep speech classification models. arXiv preprint arXiv:2408.15508 (2024)
123 Yao, W., Yang, J., He, Y., Liu, J., Wen, W.: Imperceptible rhythm backdoor attacks: Exploring rhythm transformation for embedding undetectable vulnerabilities on speech recognition. Neurocomputing **614**, 128779 (2025)
124 Ye, Z., Mao, T., Dong, L., Yan, D.: Fake the real: Backdoor attack on deep speech classification via voice conversion. In: Proceedings of the Interspeech

2023. pp. 4923–4927 (2023)
125 Ye, Z., Yan, D., Dong, L., Deng, J., Yu, S.: Stealthy backdoor attack against speaker recognition using phase-injection hidden trigger. IEEE Signal Processing Letters **30**, 1057–1061 (2023)
126 Ye, Z., Yan, D., Dong, L., Shen, K.: Breaking speaker recognition with paddingback. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4435–4439 (2024)
127 Ying, Z., Liu, A., Liu, X., Tao, D.: Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks. arXiv preprint arXiv:2406.06302 (2024)
128 Yu, Z., Chang, Y., Zhang, N., Xiao, C.: {SMACK}: Semantically meaningful adversarial audio attack. In: 32nd USENIX Security Symposium (USENIX security 23). pp. 3799–3816 (2023)
129 Yu, Z., Zhai, S., Zhang, N.: Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 460–474 (2023)
130 Yun, Z., Ao, J., Ko, T., Ronen, E., Sharif, M.: Sounding the alarm: Backdooring acoustic foundation models for physically realizable triggers (2024)
131 Ze, J., Li, X., Cheng, Y., Ji, X., Xu, W.: Ultrabd: Backdoor attack against automatic speaker verification systems via adversarial ultrasound. In: 2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS). pp. 193–200. IEEE (2023)
132 Zhang, H., Yan, Q., Zhou, P., Liu, X.Y.: Generating robust audio adversarial examples with temporal dependency. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 3167–3173 (2021)
133 Zhang, M., Ji, S., Cai, H., Dong, H., Zhang, P., Li, Y.: Audio steganography based backdoor attack for speech recognition software. In: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 1208–1217. IEEE (2024)
134 Zhang, S., Pan, Y., Liu, Q., Yan, Z., Choo, K.K.R., Wang, G.: Backdoor attacks and defenses targeting multi-domain ai models: A comprehensive review. ACM Computing Surveys **57**(4), 1–35 (2024)
135 Zhang, T., Phan, H., Tang, Z., Shi, C., Wang, Y., Yuan, B., Chen, Y.: Inaudible backdoor attack via stealthy frequency trigger injection in audio spectrogram. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. pp. 31–45 (2024)
136 Zhang, W., Zhao, S., Liu, L., Li, J., Cheng, X., Zheng, T.F., Hu, X.: Attack on practical speaker verification system using universal adversarial perturbations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2575–2579 (2021)
137 Zhang, Z., Liang, S., Shimada, D., Xu, C.: Rethinking audio-visual adversarial vulnerability from temporal and modality perspectives. In: The Thirteenth International Conference on Learning Representations (2025)
138 Zhao, S., Gan, L., Tuan, L.A., Fu, J., Lyu, L., Jia, M., Wen, J.: Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3421–3438 (2024)
139 Zhao, S., Jia, M., Guo, Z., Gan, L., XU, X., Wu, X., Fu, J., Yichao, F., Pan, F., Luu, A.T.: A survey of recent backdoor attacks and defenses in large language models. Transactions on Machine Learning Research (2025)
140 Zhao, S., Jia, M., Tuan, L.A., Pan, F., Wen, J.: Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 11507–11522 (2024)
141 Zhao, S., Tian, J., Fu, J., Chen, J., Wen, J.: Feamix: Feature mix with memory batch based on self-consistency learning for code generation and code translation. IEEE Transactions on Emerging Topics in Computational Intelligence (2024)
142 Zhao, S., Tuan, L.A., Fu, J., Wen, J., Luo, W.: Exploring clean label backdoor attacks and defense in language models. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
143 Zhao, S., Wen, J., Luu, A., Zhao, J., Fu, J.: Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12303–12317 (2023)
144 Zhao, S., Wu, X., Nguyen, C.D., Jia, Y., Jia, M., Feng, Y., Tuan, L.A.: Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation. arXiv preprint arXiv:2410.14425 (2024)
145 Zhao, S., Xu, X., Xiao, L., Wen, J., Tuan, L.A.: Clean-label backdoor attack and defense: An examination of language model vulnerability. Expert Systems with Applications **265**, 125856 (2025)
146 Zheng, B., Jiang, P., Wang, Q., Li, Q., Shen, C., Wang, C., Ge, Y., Teng, Q., Zhang, S.: Black-box adversarial attacks on commercial speech platforms with minimal information. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. pp. 86–107 (2021)
147 Zheng, Z., Li, X., Yan, C., Ji, X., Xu, W.: The silent manipulator: A practical and inaudible backdoor attack against speech recognition systems. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7849–7858 (2023)
148 Zhou, N., Lin, W., Liu, L.: Gradient norm-based fine-tuning for backdoor defense in automatic speech recognition. arXiv preprint arXiv:2502.01152 (2025)
149 Zhu, J., Chen, L., Xu, D., Zhao, W.: Backdoor defence for voice print recognition model based on speech enhancement and weight pruning. IEEE access **10**, 114016–114023 (2022)
150 Zong, W., Chow, Y.W., Susilo, W., Do, K., Venkatesh, S.: Trojanmodel: A practical trojan attack against automatic speech recognition systems. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 1667–1683. IEEE (2023)