AT-ADD: All-Type Audio Deepfake Detection Challenge Evaluation Plan

Yuankun Xie*
Communication University of China & Ant Group
Beijing, China

Haonan Cheng*
Communication University of China
Beijing, China

Jiayi Zhou*
Machine Intelligence, Ant Group
Shanghai, China

Xiaoxuan Guo
Communication University of China
Ant Group
Beijing, China

Tao Wang
Machine Intelligence, Ant Group
Shanghai, China

Jian Liu
Machine Intelligence, Ant Group
Shanghai, China

Weiqiang Wang
Machine Intelligence, Ant Group
Shanghai, China

Ruibo Fu
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Xiaopeng Wang
Beijing Institute of Technology
Beijing, China

Hengyan Huang
Communication University of China
Beijing, China

Xiaoying Huang
Communication University of China
Beijing, China

Long Ye
Communication University of China
Beijing, China

Guangtao Zhai
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

The rapid advancement of Audio Large Language Models (ALLMs) has enabled cost-effective, high-fidelity generation and manipulation of both speech and non-speech audio, including sound effects, singing voices, and music. While these capabilities foster creativity and content production, they also introduce significant security and trust challenges, as realistic audio deepfakes can now be generated and disseminated at scale. Existing audio deepfake detection (ADD) countermeasures (CMs) and benchmarks, however, remain largely speech-centric, often relying on speech-specific artifacts and exhibiting limited robustness to real-world distortions, as well as restricted generalization to heterogeneous audio types and emerging spoofing techniques. To address these gaps, we propose the *All-Type Audio Deepfake Detection (AT-ADD)* Grand Challenge for ACM Multimedia 2026, designed to bridge controlled academic evaluation with practical multimedia forensics. AT-ADD comprises two tracks: (1) *Robust Speech Deepfake Detection*, which evaluates detectors under real-world scenarios and against unseen, state-of-the-art speech generation methods; and (2) *All-Type Audio Deepfake Detection*, which extends detection beyond speech to diverse, unknown audio types and promotes type-agnostic generalization across speech, sound, singing, and music. By providing standardized datasets, rigorous evaluation protocols, and reproducible baselines, AT-ADD aims to accelerate the development of robust and generalizable audio forensic technologies, supporting secure communication, reliable media verification, and responsible governance in an era of pervasive synthetic audio.

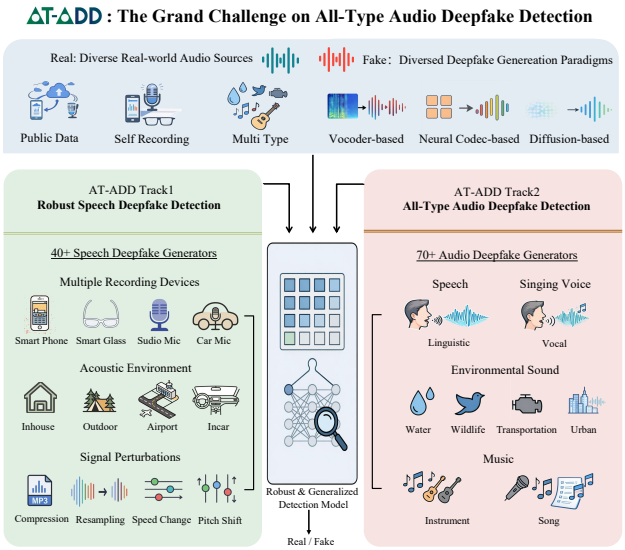

Figure 1: AT-ADD challenge overview.

KEYWORDS

Audio Deepfake Detection, Countermeasure, Audio Large Language Model

1 INTRODUCTION

Recent advances in audio generation technologies, particularly Audio Large Language Models (ALLMs), have significantly improved the realism, scalability, and accessibility of synthetic audio. Modern generative systems are now capable of producing high-fidelity audio across a wide range of content types, including speech, environmental sounds, singing voices, and music. While these developments

*These authors contributed equally to this work.
Official Website: https://at-add.com.
Organization Inquiries: Haonan Cheng (haonancheng@cuc.edu.cn), Jiayi Zhou (zjy326112@antgroup.com).
Technical Questions: Yuankun Xie (xieyuankun@cuc.edu.cn), Tao Wang (mengyu.wt@antgroup.com).

greatly benefit content creation and multimedia applications, they also introduce serious security and trust risks, as audio deepfakes can be generated and disseminated at scale with increasing realism.

Despite growing research efforts in audio deepfake detection (ADD), existing methods and benchmarks remain largely focused on speech and are typically evaluated under relatively controlled conditions. Consequently, current countermeasures (CMs) often rely on speech-specific artifacts and exhibit limited robustness when deployed in real-world scenarios involving channel variability, environmental noise, compression, replay attack and other distortions. Furthermore, their generalization capability remains insufficient when faced with emerging generation paradigms, such as ALLM-based synthesis and neural codec-driven generation, as well as diverse non-speech audio types.

To address these challenges, we introduce the AT-ADD (All-Type Audio Deepfake Detection) Grand Challenge at ACM Multimedia 2026. The goal of AT-ADD is to bridge the gap between idealized research settings and real-world multimedia forensics by systematically evaluating both robustness under realistic conditions and generalization across audio types and unseen generation methods.

Track 1: Robust Speech Deepfake Detection. This track focuses on robustness in real-world speech deepfake detection. We construct a large-scale dataset (AT-ADD Track 1) covering more than 40 state-of-the-art speech generation models, spanning vocoder-based, neural codec-based, and diffusion-based paradigms, with particular emphasis on emerging ALLM-driven synthesis. The real speech data are collected from multiple public datasets and in-the-wild recordings, covering diverse languages, recording devices (e.g., smartphones, in-vehicle systems, wearable devices), and acoustic environments. To simulate realistic deployment conditions, we further introduce a wide range of degradations, including background noise, reverberation, replay attack, compression, resampling, and speed perturbation. This track aims to evaluate the robustness and cross-domain generalization of detection systems under complex real-world conditions, encouraging models to move beyond reliance on narrow artifact cues toward learning more fundamental and transferable forensic representations.

Track 2: All-Type Audio Deepfake Detection. Building upon Track 1, this track extends the detection task from speech to all types of audio, including environmental sounds, singing voices, and music. We construct a new dataset (AT-ADD Track 2) covering more than 70 audio generation models across different generation mechanisms. Unlike Track 1, no additional signal-level perturbations are introduced in this track, allowing for a more controlled investigation of cross-type generalization. Real audio is collected from multiple public datasets, while synthetic audio is generated under a unified framework without introducing additional distortions. Participants are required to design type-agnostic CMs capable of distinguishing real and fake audio under unseen audio categories. This track encourages models to capture shared synthesis artifacts across different audio types, promoting the development of universal audio deepfake detection methods.

Together, the two tracks form a progressive evaluation framework: Track 1 emphasizes robustness under realistic conditions within the speech domain, while Track 2 focuses on generalization across audio types and domains. This design reflects the evolving landscape of audio deepfake threats—from speech-centric manipulation to open-domain audio generation—and provides a structured benchmark for advancing robust and generalizable audio forensic technologies.

2 CHALLENGE TASKS

The AT-ADD challenge is organized into two complementary and progressively structured tracks under a **closed setting**. This design aims to assess which types of CMs can achieve stronger generalization and robustness under limited data conditions. Participants are required to train their CMs strictly using only the data provided by the organizers, thereby ensuring a fair and controlled comparison across methods.

2.1 Track 1: Robust Speech Deepfake Detection

Goal. Track 1 aims to bridge the gap between existing benchmarks and real-world deployment scenarios for speech deepfake detection. It evaluates whether a detector can remain reliable under realistic domain shifts and practical post-processing effects, while maintaining strong performance against modern high-fidelity synthesis systems.

Task definition. Given an input speech utterance, participants are required to predict whether the input is *real* or *fake*. In this task, *fake* refers specifically to deepfake speech generated using deep neural network-based methods, while *real* refers to non-deepfake speech. It should be noted that signal distortions or transformations, such as compression, resampling, speed perturbation, and pitch shifting, as well as replay-based attacks, do not change the original real/fake label in this task. The training and development data are fully provided by the organizers, and the use of external data is not allowed under the closed setting.

The evaluation set includes deepfake samples generated by methods that are *unseen* during training and reflect recent state-of-the-art generation techniques. Meanwhile, the real speech in the evaluation set is collected under realistic conditions, involving variations in recording devices, acoustic environments, languages, and other real-world factors.

Long-term research target. Track 1 is designed to promote the development of practically deployable CMs that generalize across unseen generators and real-world recording conditions. By combining controlled training data with an evaluation set featuring both unseen deepfake methods and significant domain shifts, this track encourages advances in robustness, generalization, and reliability beyond conventional clean-benchmark settings.

2.2 Track 2: All-Type Audio Deepfake Detection

Goal. Track 2 targets universal audio deepfake detection across heterogeneous audio types and aims to develop *type-agnostic* detectors that generalize across both audio types and unseen generation methods.

Task definition. Given an input audio clip of unknown type, participants are required to determine whether it is *real* or *fake*. In this task, *fake* denotes deepfake audio generated by deep neural network-based methods, whereas *real* denotes non-deepfake audio. Notably, in Track 2, audio-type labels (i.e., speech, sound, singing, and music) are *not* available at test time, reflecting realistic deployment scenarios.

Similar to Track 1, this track follows a closed setting, where participants must use only the provided training and development data, without access to external resources.

Long-term research target. Track 2 is designed to move beyond speech-centric audio deepfake detection and promote the development of unified CMs capable of handling both unknown audio types and unseen generation methods. This setting is motivated by real-world scenarios, where the type of incoming audio is often unknown in advance and may extend beyond speech alone. By encouraging the learning of shared and transferable representations across diverse audio types, Track 2 aims to advance universal CMs that are better suited to realistic and heterogeneous application environments.

3 RELATED WORK

ADD has progressed rapidly with the rise of high-fidelity neural generative models. Existing studies are dominated by speech-focused benchmarks and methods, where SSL representations and attention-based back-ends have driven substantial performance gains. In contrast, detection for non-speech audio (sound, singing voice, and music) remains less explored and is still largely benchmark-driven, while cross-type generalization across heterogeneous audio domains is an emerging yet under-established research frontier. In the following, we review prior work by audio type—speech, sound, singing voice, and music—and then discuss cross-type ADD, which motivates the need for a unified and realistic all-type benchmark.

Speech. Speech deepfake detection has been extensively studied, largely driven by the ASVspoof challenges [38, 40, 57] and ADD challenges [66, 67]. Representative CMs include AASIST [20] and SSL-based pipelines that combine XLSR with AASIST [56]. Subsequent studies have investigated different SSL representations [22, 45], layer utilization of SSL features [43, 59, 70], and robustness [21, 63, 75]. However, a substantial gap remains between existing public benchmarks and real-world conditions (e.g., diverse capture devices, channel effects, and replay attack [39, 72]), highlighting the need for new datasets and protocols that better reflect realistic deployment scenarios and enable more faithful evaluation of detection performance in the wild.

Sound. Compared to speech deepfake detection, research on environmental sound deepfake detection is still relatively nascent and is largely driven by dataset and benchmark construction. The Environmental Sound Deepfake Detection (ESDD) Challenge [68] has recently advanced this area by covering a wide range of ALLM-based text-to-audio (TTA) and audio-to-audio (ATA) synthesis methods. Current state-of-the-art solutions typically leverage sound-oriented SSL representations such as SSLAM [13].

Singing voice. Singing voice can be considered a subcategory of music; however, it is treated as a distinct audio type in this challenge due to its unique characteristics and the high difficulty of deepfake song detection. Unlike general music, singing voice shares strong similarities with speech as both are produced by human vocal mechanisms, while also exhibiting complex musical structures such as melody and rhythm. These properties make singing voice particularly challenging for existing CMs. Recent work, such as SVDD [74], has promoted research in deepfake singing voice detection. Competitive approaches often leverage hybrid representations by combining speech-oriented SSL features (e.g., XLSR) with music-oriented SSL models such as MERT [32] and WavLM [5], as explored in recent studies [6, 15, 71].

Music. Music represents a broad and diverse audio type that encompasses both instrumental compositions and songs. Compared to speech and singing voice, music exhibits higher variability in structure, timbre, and generation mechanisms, posing additional challenges for deepfake detection. FakeMusicCaps [7] provides a benchmark for synthetic-music detection and enables the study of text-to-music (TTM) generation artifacts. However, methodological explorations remain relatively limited compared to speech [31, 60].

Cross-type. A few studies have investigated transfer across audio types, for example between speech and singing voice [12], from speech to music [30], and ESDD 2 (from sound to both speech and sound deepfake detection settings [73]). Xie et al. [61] further establish an SSL-based benchmark for all-type ADD and propose wavelet prompt tuning to improve cross-type generalization. However, these studies are still grounded on relatively limited and task-specific datasets, and the community is still lacking a comprehensive and widely accepted benchmark that systematically covers the full spectrum of audio types and realistic conditions.

Overall, despite substantial progress in speech deepfake detection, a clear gap remains between academic benchmarks and real-world deployment, particularly in terms of robustness to complex acoustic environments, and rapidly evolving deepfake paradigms. This motivates the need for a robust speech deepfake detection benchmark that more faithfully reflects practical scenarios. Meanwhile, research on all-type audio deepfake detection is still at an early stage, and a unified evaluation model spanning speech, sound, singing voice, and music is essential to drive the next generation of generalizable CMs.

4 DATASETS AND RESOURCES

To support the AT-ADD challenge, we construct two benchmark datasets: **AT-ADD Track 1** and **AT-ADD Track 2**, corresponding to robust speech deepfake detection and all-type audio deepfake detection, respectively. For both tracks, we provide standardized train, development (dev), and evaluation (eval) splits under a closed setting. In addition, a progress subset is provided for progress evaluation, which is sampled from the evaluation set with the same distribution and constitutes 20% of the full eval set. An overview of the two tracks is presented in Table 1.

Detailed dataset compositions and statistics are presented in the following subsections. It should be noted that, to ensure data quality (e.g., by removing fully silent segments), we applied a series of screening and filtering procedures. As a result, the number of samples in each condition is not perfectly uniform.

4.1 Track 1: Robust Speech Deepfake Detection

We construct the AT-ADD Track 1 dataset with predefined training, development, and evaluation splits. The evaluation split is reserved for testing and consists of real speech collected from diverse domains, together with fake speech generated by methods that are unseen in the training and development sets, thereby enabling the evaluation of CM robustness under domain shifts, such as variations in recording devices, acoustic environments, and signal perturbations. Table 2 summarizes the overall composition of AT-ADD Track 1 dataset.

Table 1: AT-ADD statistics (number of clips) for Track 1 and Track 2. * indicates withheld statistics.

Split	T1	T2				
		Speech	Sound	Singing	Music	Total
Train	49,575	49,575	39,840	36,000	21,366	146,781
Dev	49,734	49,734	19,929	16,000	5,406	91,069
Progress	29,269	*	*	*	*	45,875
Eval	146,346	*	*	*	*	229,373

Table 2: Details of our proposed AT-ADD Track 1 dataset.

Model / Condition	Train	Dev
Real	9,999	10,000
ProDiff [18]	1,999	1,998
PortaSpeech (normal) [50]	1,996	1,994
DiffSpeech [36]	1,998	1,999
FastSpeech2 [49]	1,998	1,998
Kokoro [41]	2,000	1,998
WaveNet [42]	1,999	2,000
FastDiff [17]	1,996	1,994
MeloTTS [76]	1,996	1,988
CosyVoice [10]	1,835	2,000
Parler-TTS (mini) [28]	1,998	1,995
GradTTS [46]	1,994	1,997
FastPitch [29]	1,998	1,997
Tacotron2 [52]	1,994	1,998
Glow-TTS [24]	2,000	2,000
WaveGlow [47]	1,997	1,997
MultiBandMelGAN [27]	1,997	2,000
Tortoise-TTS [3]	1,992	1,989
StarGANv2-VC [33]	1,999	1,999
Llasa 1B [65]	1,793	1,795
Index-TTS [9]	1,997	1,998
Total	**49,575**	**49,734**

Real speech in train/dev sets. The real speech subset within the training and development sets is comprised of a diverse collection of multilingual utterances. This subset incorporates internal recordings captured across a variety of Recording devices to ensure acoustic diversity, alongside high-quality Chinese speech samples sourced from the AISHELL-3 [54] dataset. To bolster the English portion, samples are integrated from the LibriTTS-R [25] and LJSpeech [19] corpora. Furthermore, the dataset's multilingual breadth is further extended through the inclusion of representative samples from Common Voice [2], covering a wide array of linguistic contexts.

Fake speech in the train/dev sets. The fake speech subset covers multiple tasks, including text-to-speech (TTS) and voice conversion (VC). For TTS, the input texts are selected from the real speech data described in the previous subsection and used for synthesis. For VC and one-shot TTS tasks that require reference speaker cloning, we use the same pool of real reference speakers for the training and development sets, while a different pool of real reference speakers is used for the evaluation set to prevent speaker-information leakage. In addition, during cloning, we ensure that the reference speaker and the source speaker are never the same person.

Table 3: AT-ADD Track 2 subsets: `train/dev` composition across audio types.

Type	Model / Source	Train	Dev
Sound	Real	9,854	4,940
	AudioLDM [34]	9,995	4,997
	AudioLDM 2 [35]	10,000	5,000
	AudioGen [26]	9,991	4,992
	Total	**39,840**	**19,929**
Singing	Real	9,000	4,000
	Soft-VITS-SVC [55]	9,000	4,000
	NeuCoSVC [51]	9,000	4,000
	SeedVC [37]	9,000	4,000
	Total	**36,000**	**16,000**
Music	Real	4,297	536
	MusicGen [8]	4,212	1,204
	MusicLDM [4]	4,276	1,221
	AudioLDM2 [35]	4,289	1,222
	Stable Audio Open [11]	4,292	1,223
	Total	**21,366**	**5,406**

Eval sets. The real speech in the evaluation set is collected from both self-recorded data captured with diverse recording devices and additional out-of-domain (OOD) public-source data. The fake speech in the evaluation set is generated by 26 methods that are *unseen* during training. In terms of synthesis paradigms, the fake speech covers several mainstream categories, including vocoder-based, codec-based, and diffusion-based methods. Furthermore, a portion of the fake speech is further replayed to simulate replay attack scenarios, while another portion of both real and fake speech is subjected to signal perturbations to thoroughly evaluate countermeasure robustness. Importantly, neither replay nor signal perturbation changes the original real/fake label of an audio sample; rather, they are regarded as markers for evaluating CM robustness.

4.2 Track 2: All-Type Audio Deepfake Detection

In this section, we describe the composition of the proposed AT-ADD Track 2 dataset across four different audio types. Table 3 summarizes the overall composition of AT-ADD Track 2 dataset.

Speech. We use the same speech training and development sets as in AT-ADD Track 1. However, the evaluation set is simplified by removing the signal perturbation and replay attak introduced in Track 1. As Track 2 targets universal, all-type deepfake detection, this design provides a clean and consistent evaluation protocol that emphasizes cross-type generalization rather than robustness to signal degradations.

Sound. The sound subset is constructed from the AudioCaps dataset [23]. We first divide the audio samples in AudioCaps, which labeled as real samples, into non-overlapping training, development, and evaluation sets. The synthetic samples in the training and development sets are generated by TTA models conditioned on the corresponding textual descriptions of the real audio. For the evaluation set, the fake samples are generated from the remaining textual descriptions using 4 *unseen* generation methods. In addition, we include OOD real sound samples from other public datasets in the evaluation set to assess the generalization capability of CMs.

Singing Voice. The singing voice subset is constructed from three source datasets: OpenCpop [58], M4Singer [69], and KiSing [53]. As in the sound subset, the samples from these three source domains are first divided into non-overlapping training, development, and evaluation sets, which are labeled as the real samples. The fake samples in the training and development sets are generated via singing voice conversion, with strictly non-overlapping source and target singers to avoid identity leakage. The fake samples in the evaluation set are produced by 5 *unseen* deepfake methods, allowing a comprehensive evaluation of cross-model generalization.

Music. The music subset is derived from the MusicCaps [1] dataset. We first divide the audio samples in MusicCaps, which labeled as real samples, into non-overlapping training, development, and evaluation sets. The synthetic samples in the training and development sets are generated by TTM models conditioned on the corresponding textual descriptions of the real music. For the evaluation set, the fake samples are generated from the remaining textual descriptions using 4 *unseen* generation methods. In addition, we include OOD real music samples from other public datasets in the evaluation set to assess the generalization capability of CMs.

5 BASELINES

To facilitate fair comparison and lower the entry barrier, we provide a set of official baselines covering conventional CMs, SSL-based CMs, and ALLM-based CMs. These baselines span different modeling paradigms and serve as strong and reproducible starting points for participants.

5.1 Baseline Models

We provide official implementations for all baseline systems used in AT-ADD, including conventional and SSL-based baselines[1] as well as the ALLM-based baseline[2]. We next give a brief introduction to these baseline models.

Conventional CMs. These models follow the traditional pipeline of feature extraction and discriminative classification, representing standard approaches in audio deepfake detection. Although generally weaker than recent SSL-based methods, they serve as important reference systems for evaluating robustness and cross-type generalization.

- **Spec-ResNet**: A spectrogram-based detector with a ResNet backbone [16], representing traditional feature-based CMs. In this baseline, we use only STFT-based spectrograms as input features, without incorporating specialized speech features such as Mel-spectrograms. This design aims to investigate whether simple, generic spectral representations can provide robustness to noise and support cross-type generalization.
- **AASIST [20]**: A raw waveform-based model employing a sinc convolution front-end [48] and residual blocks, followed by spectral–temporal attention for classification. This baseline operates directly on raw waveforms without explicit feature engineering, allowing us to study the performance of end-to-end waveform-based detection.

SSL-based CMs. To improve robustness and generalization, we further include models enhanced by SSL representations. These approaches leverage large-scale pre-training and have become the dominant paradigm in modern CM systems, demonstrating strong performance.

- **FT-XLSR-AASIST [56]**: An enhanced AASIST model using self-supervised representations Wav2Vec2-XLSR[3] as the front-end, providing improved transferability across languages and domains [44, 45]. FT denotes full fine-tuning (FT) of all layers in XLSR and AASIST. To date, it remains a competitive baseline for audio deepfake detection.
- **WPT-XLSR-AASIST [61]**: A strengthened SSL-based baseline incorporating wavelet prompt tuning (WPT) to capture frequency-invariant artifacts. By only optimizing a small number of prompt tokens, this method achieves strong performance with minimal training cost.

ALLM-based CMs. We further provide ALLM-based CMs built on the Qwen audio family.These models take audio (optionally with textual instructions) as input and generate textual outputs, which are adapted into binary real/fake predictions via supervised fine-tuning (SFT). Compared to conventional CMs, ALLM-based approaches leverage large-scale multimodal pre-training and exhibit strong potential for cross-type generalization and unified modeling across heterogeneous audio domains.

- **Qwen2.5-Omni-3B / 7B**[4][5]: Unified multimodal models that support audio inputs and can be adapted for audio deepfake detection via supervised fine-tuning (SFT). Compared to conventional CMs, ALLM-based CMs are capable of producing deterministic predictions and can further provide interpretable reasoning through techniques such as reinforcement learning. Recent studies have demonstrated the superior performance of ALLM-based approaches in the field of ADD [14, 62, 64]. Their potential for improving robustness and enabling unified modeling across all audio types makes them a promising direction for further exploration in this challenge.

5.2 Baseline Performance

Table 4 reports the performance of all baselines on both Track 1 and Track 2 under the official evaluation metrics. Notably, for conventional and SSL-based CMs, a unified decision threshold of 0.5 is adopted for real/fake classification.

SSL-based CMs achieve the strongest performance on both tracks. In particular, FT-XLSR-AASIST reaches the best results with 76.73% on Track 1 (Eval) and 79.47% on Track 2 (Eval), consistently outperforming all other baselines. Its performance is also stable across

[1] https://github.com/xieyuankun/AT-ADD-Baseline
[2] https://github.com/yangchunmian123/AT-ADD-ALLM-Baseline
[3] https://huggingface.co/facebook/wav2vec2-xls-r-300m
[4] https://huggingface.co/Qwen/Qwen2.5-Omni-3B
[5] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

Table 4: Baseline performance (%) on the AT-ADD benchmarks. Best results are highlighted in bold.

Type	Model	T1			T2			T2 Progress (by Type)				T2 Eval (by Type)			
		Dev	Progress	Eval	Dev	Progress	Eval	Speech	Sound	Singing	Music	Speech	Sound	Singing	Music
Conventional	Spec-ResNet	81.85	47.93	47.41	57.51	53.22	53.83	51.08	52.92	48.41	60.48	51.79	54.35	49.29	59.88
	AASIST	94.16	60.78	60.39	93.63	62.38	62.21	63.69	56.88	64.08	64.87	63.58	56.62	63.81	64.85
SSL-based	FT-XLSR-AASIST	**99.70**	**76.98**	**76.73**	**98.48**	**79.25**	**79.47**	**79.43**	**66.08**	**96.33**	**75.17**	**79.50**	**66.82**	**96.30**	**75.28**
	WPT-XLSR-AASIST	96.27	73.56	73.35	95.00	66.59	66.68	69.42	52.97	79.81	64.17	69.31	53.83	79.56	64.04
ALLM-based	Qwen2.5-Omni-3B	93.97	68.65	68.02	94.44	63.42	63.23	69.47	50.38	66.31	67.52	68.74	50.41	65.78	67.99
	Qwen2.5-Omni-7B	95.93	69.19	68.64	94.70	61.48	61.78	69.89	45.28	68.04	63.77	69.29	45.94	68.03	63.87

audio types, achieving 79.50% (speech), 66.82% (sound), 96.30% (singing), and 75.28% (music) on the Track 2 evaluation set.

Conventional CMs show significantly lower performance. For example, Spec-ResNet achieves only 47.41% on Track 1 (Eval) and 53.83% on Track 2 (Eval), while AASIST improves to 60.39% and 62.21%, respectively, but still remains substantially below SSL-based approaches.

ALLM-based models demonstrate competitive performance despite their general-purpose design. Qwen2.5-Omni-7B achieves 68.64% on Track 1 (Eval) and 61.78% on Track 2 (Eval), while the 3B variant achieves comparable results (68.02% / 63.23%). Notably, ALLM-based models exhibit relatively balanced performance across audio types, suggesting their potential for unified modeling in the all-type ADD setting.

Comparing the two tracks, the overall difficulty of Track 1 and Track 2 is comparable from the baseline results. Track 1 mainly challenges the robustness of models to unseen generators and real-world variations in real audio, while Track 2 highlights performance gaps across audio types (e.g., lower scores on sound and music compared to singing), indicating remaining challenges in achieving fully uniform cross-type generalization.

6 RULES

This section outlines the evaluation metrics, and rules for both tracks. Participants are expected to adhere to these rules to ensure a fair and transparent competition.

6.1 Evaluation Metrics

The performance of submitted systems is evaluated using the $F1$-score. To ensure fair comparison under class imbalance and across audio types, Macro-$F1$ is adopted with different aggregation strategies for the two tracks.

For a given class c, the $F1$-score is defined as the harmonic mean of precision P_c and recall R_c:

$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{1}$$

where

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

and TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively.

Track 1. For Track 1 (binary classification: *real* vs. *fake*), the official metric is the Macro-$F1$ over the two classes:

$$\text{Macro-}F1_{\text{T1}} = \frac{1}{2}\left(F1_{\text{real}} + F1_{\text{fake}}\right) \tag{3}$$

This metric assigns equal importance to both classes, making it robust to class imbalance.

Track 2. For Track 2, the evaluation accounts for both class balance and audio-type balance. Specifically, for each audio type t, we first compute a type-wise Macro-$F1$:

$$\text{Macro-}F1_t = \frac{1}{2}\left(F1_{t,\text{real}} + F1_{t,\text{fake}}\right) \tag{4}$$

The final score is then obtained by averaging over all audio types:

$$\text{Macro-}F1_{\text{T2}} = \frac{1}{4}\sum_{t=1}^{4} \text{Macro-}F1_t \tag{5}$$

where $t \in \{\text{speech}, \text{sound}, \text{singing}, \text{music}\}$.

Thus, the Track 2 metric enforces a two-level balance: equal weighting across audio types and equal weighting between *real* and *fake* classes within each type.

6.2 Competition Rules

The following rules apply to both Track 1 and Track 2 unless otherwise specified.

Data Usage. Participants may use only the officially released training and development sets for model training, validation, model selection, and threshold determination. They are free to split the released data for internal training and validation, and may also merge the training and development sets for training. The progress and evaluation sets must not be used in any form for training, fine-tuning, pseudo-labeling, self-training, threshold tuning, or any other kind of model adaptation.

External Data. Except for the officially released data, the use of any external labeled or unlabeled audio data is strictly prohibited for training, fine-tuning, distillation, calibration, or pseudo-label construction, including self-generated synthetic data from external generative models or services. Participants must not introduce external datasets related to audio deepfake detection or other closely related authenticity-discrimination tasks, nor may they use models, checkpoints, or feature extractors that have been pre-trained, trained, or fine-tuned on such datasets.

Data Augmentation. Data augmentation is allowed only in the form of signal-level perturbation or transformation applied to the officially released data, rather than by introducing external audio data as additional training samples. Allowed augmentation strategies include, but are not limited to, additive noise, reverberation, compression, resampling, and signal-level augmentation methods such as RawBoost. Publicly available augmentation resources, such

as MUSAN and RIR libraries, may be used only as augmentation sources and must not be treated as additional supervised training data.

Pretrained Models. Publicly available and traceable pretrained models are allowed, including self-supervised learning (SSL) models, audio large language models (ALLMs), multimodal large language models (MLLMs), and other general-purpose pretrained models, provided that their sources can be clearly specified in the final metadata. However, any external models, checkpoints, or feature extractors that have been supervisedly trained or fine-tuned outside AT-ADD for audio deepfake detection or other closely related authenticity classification tasks are strictly prohibited.

Fusion and Ensemble. Fusion and ensemble strategies are allowed, including feature-level fusion, score-level fusion, and decision-level fusion. The final submitted system may contain no more than 5 subsystems, and all components must comply with the same data usage rules.

Reproducibility. The system corresponding to the final submitted score must be fully automatic and reproducible. The use of opaque closed-source APIs or any other external services that cannot be independently reproduced by the organizers is prohibited. Manual intervention in test set prediction, listening-based correction, or manual annotation is not allowed.

Compliance Check. For top-ranked teams, the organizers reserve the right to request a method description, a resource declaration, a model list, and inference code. If any violation of the data usage rules is found, or if a system is determined to be non-reproducible or to involve test-set leakage, the organizers reserve the right to disqualify the submission.

7 PARTICIPATION INSTRUCTIONS

To participate in AT-ADD, teams must first request dataset access by completing the registration form on the Hugging Face dataset page for Track 1[6] and Track 2[7]. By participating in the challenge, participants are deemed to have agreed to the corresponding *AT-ADD-Dataset-License*.

After obtaining dataset access, participants are required to register on Codabench for evaluation on Track 1[8] and Track 2[9]. Each team is allowed to use only **one** Codabench account, and the email address used for Codabench registration must be consistent with that used during Hugging Face registration.

For submission, participants must upload a `.zip` file, whose filename can be arbitrary. The compressed file must contain a single file named `predict.csv`. The required format is:

```
name,predict
ATADD_T1_Eval_000001.flac,fake
ATADD_T1_Eval_000002.flac,real
...
```

Here, `name` denotes the audio filename, and `predict` indicates the predicted label, which must be either `real` or `fake`.

[6]https://huggingface.co/datasets/xieyuankun/AT-ADD-Track1
[7]https://huggingface.co/datasets/xieyuankun/AT-ADD-Track2
[8]https://www.codabench.org/competitions/15477
[9]https://www.codabench.org/competitions/15481

8 CHALLENGE SCHEDULE

The AT-ADD challenge schedule is strictly aligned with the ACM Multimedia 2026 Grand Challenge timeline. The competition consists of a development phase, a final evaluation period, and a technical reporting stage.

Table 5: AT-ADD Challenge Important Dates

Event	Date (2026)
Release of Data, Baseline Code, and Paper	April 8
Opening of Progress Evaluation Stage	April 8
Opening of Final Evaluation Stage	June 8
Final Leaderboard Freeze	June 15
Metadata and Technical Report Submission	June 17
Participant Paper Submission Deadline	**June 25**
Notification of Paper Acceptance	July 16
Camera-Ready Paper Submission Deadline	**August 6**

The challenge schedule is organized into the following key stages:

- **Progress Evaluation Phase (April 8 – June 11):** Upon the official release of the dataset, baseline code, and challenge paper on April 8, the Progress evaluation stage for all tracks opens on April 8. During this phase, participants can utilize the provided training and development sets to optimize their models and receive real-time feedback via the preliminary leaderboard.
- **Final Evaluation Phase (June 8–June 15):** The final evaluation phase begins on June 8, during which participants in all tracks are required to submit predictions on the full evaluation set. The final leaderboard will be frozen on June 15. The final evaluation dataset will be announced to all registered participants on June 8 via the Codabench mailing system.
- **Metadata and Technical Report Submission (June 15–June 17):** After the leaderboard is frozen, each team is required to submit a metadata form describing its submitted system. The detailed format and submission instructions will be sent by email to all teams that have participated in the evaluation phase, and the submission deadline is June 17. In addition to the mandatory metadata, teams may also submit optional technical reports to present their methods in greater detail. These reports will also be displayed on the AT-ADD official website. Although optional, such reports are strongly encouraged, as they help the organizers better understand the proposed methods. Based on the final rankings and submitted materials, three awards will be presented: first place in Track 1, first place in Track 2, and the Best Solution Award. The Best Solution Award will be determined by an expert panel based on algorithmic innovation, practical applicability, interpretability, and reusability.
- **Paper Submission and Review (June 25 – July 16):** The three award-winning teams will be invited to prepare and submit their papers to the official ACM MM submission system by June 25. These papers will undergo the standard peer-review process of ACM Multimedia, with acceptance notifications scheduled for July 16.
- **Final Camera-Ready Submission (August 6):** Authors of accepted papers are required to submit their final camera-ready versions by August 6 for inclusion in the main conference proceedings.

REFERENCES

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*. 4218–4222.

[3] James Betker. 2023. Better Speech Synthesis Through Scaling. *arXiv preprint arXiv:2305.07243* (2023).

[4] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1206–1210.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[6] Xuanjun Chen, Haibin Wu, Roger Jang, and Hung-yi Lee. 2024. Singing Voice Graph Modeling for SingFake Detection. In *Proc. Interspeech 2024*. 4843–4847.

[7] Luca Comanducci, Paolo Bestagini, and Stefano Tubaro. 2024. Fakemusiccaps: a dataset for detection and attribution of synthetic music generated via text-to-music models. *arXiv preprint arXiv:2409.10684* (2024).

[8] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. 2025. IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-to-Speech System. *arXiv preprint arXiv:2502.05512* (2025).

[10] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, et al. 2024. CosyVoice: A Scalable Multilingual Zero-Shot Text-to-Speech Synthesizer Based on Supervised Semantic Tokens. *arXiv preprint arXiv:2407.05407* (2024).

[11] Zach Evans, Julian D. Parker, C. J. Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable Audio Open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[12] Mahyar Gohari, Davide Salvi, Paolo Bestagini, and Nicola Adami. 2025. Audio Features Investigation for Singing Voice Deepfake Detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[13] Xiaoxuan Guo, Hengyan Huang, Jiayi Zhou, Renhe Sun, Jian Liu, Haonan Cheng, Long Ye, and Qin Zhang. 2025. EnvSSLAM-FFN: Lightweight Layer-Fused System for ESDD 2026 Challenge. *arXiv preprint arXiv:2512.20369* (2025).

[14] Xiaoxuan Guo, Yuankun Xie, Haonan Cheng, Jiayi Zhou, Jian Liu, Hengyan Huang, Long Ye, and Qin Zhang. 2026. Towards Explicit Acoustic Evidence Perception in Audio LLMs for Speech Deepfake Detection. *arXiv preprint arXiv:2601.23066* (2026).

[15] Anmol Guragain, Tianchi Liu, Zihan Pan, Hardik B. Sailor, and Qiongqiong Wang. 2024. Speech Foundation Model Ensembles for the Controlled Singing Voice Deepfake Detection (CTRSVDD) Challenge 2024. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. 774–781. doi:10.1109/SLT61566.2024.10832226

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Rongjie Huang, Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. *arXiv preprint arXiv:2204.09934* (2022).

[18] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of ACM MM*. 2595–2605.

[19] K. Ito and L. Johnson. 2017. The LJ speech dataset. *https://keithito. com/LJ-Speech-Dataset/* (2017).

[20] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of the ICASSP*. 6367–6371.

[21] Piotr Kawa, Marcin Plata, and Piotr Syga. 2023. Defense Against Adversarial Attacks on Audio DeepFake Detection. In *Proc. Interspeech 2023*. 5276–5280.

[22] Yassine El Kheir, Youness Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. 2025. Comprehensive Layer-wise Analysis of SSL Models for Audio Deepfake Detection. *arXiv preprint arXiv:2502.03559* (2025).

[23] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 119–132.

[24] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems* 33 (2020), 8067–8077.

[25] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802* (2023).

[26] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. AudioGen: Textually Guided Audio Generation. In *The Eleventh International Conference on Learning Representations*.

[27] K. Kumar, Thibault Kumar, R., L. Gestin, W. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).

[28] Yann Lacombe, Vaibhav Srivastav, and Sahaj Gandhi. 2024. Parler-TTS. https://github.com/huggingface/parler-tts.

[29] Adrian Łańcucki. 2021. FastPitch: Parallel Text-to-Speech with Pitch Prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6588–6592.

[30] Yupei Li, Manuel Milling, Lucia Specia, and Björn W Schuller. 2024. From Audio Deepfake Detection to AI-Generated Music Detection–A Pathway and Overview. *arXiv preprint arXiv:2412.00571* (2024).

[31] Yupei Li, Qiyang Sun, Hanqian Li, Lucia Specia, and Björn W Schuller. 2024. Detecting Machine-Generated Music with Explainability–A Challenge and Early Benchmarks. *arXiv preprint arXiv:2412.13421* (2024).

[32] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. 2024. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In *ICLR*.

[33] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. 2021. StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion. *arXiv preprint arXiv:2107.10394* (2021).

[34] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *arXiv preprint arXiv:2301.12503* (2023).

[35] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation with Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 2871–2883.

[36] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 11020–11028.

[37] Songting Liu. 2024. Zero-Shot Voice Conversion with Diffusion Transformers. *arXiv preprint arXiv:2411.09943* (2024).

[38] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

[39] Nicolas Müller, Piotr Kawa, Wei-Herng Choong, Adriana Stan, Aditya Tirumala Bukkapatnam, Karla Pizzi, Alexander Wagner, and Philip Sperl. 2025. Replay Attacks Against Audio Deepfake Detection. In *Proc. Interspeech 2025*. 2245–2249.

[40] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. 2021. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 2 (2021), 252–265.

[41] Aryan Nayak. 2025. Kokoro: An Accessible Text-to-Speech Application for Visually Impaired Students. Independent publication.

[42] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[43] Zihan Pan, Tianchi Liu, Hardik B Sailor, and Qiongqiong Wang. 2024. Attentive Merging of Hidden Embeddings from Pre-trained Speech Model for Anti-spoofing Detection. In *Proc. Interspeech 2024*. 2090–2094.

[44] Octavian Pascu, Adriana Stan, Dan Oneata, Elisabeta Oneata, and Horia Cucu. 2024. Towards generalisable and calibrated audio deepfake detection with self-supervised representations. In *Interspeech 2024*. 4828–4832. doi:10.21437/Interspeech.2024-1302

[45] Orchid Chetia Phukan, Gautam Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024. Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 2496–2506.

[46] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 8599–8608.

[47] R. Prenger, R. Valle, and B. Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3617–3621.

[48] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 1021–1028.

[49] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*. https://openreview.net/forum?id=piLPYqxtWuA

[50] Yi Ren, Jinglin Liu, and Zhou Zhao. 2021. PortaSpeech: Portable and High-Quality Generative Text-to-Speech. In *Advances in Neural Information Processing Systems*, Vol. 34. 13963–13974.

[51] Binzhu Sha, Xu Li, Zhiyong Wu, Ying Shan, and Helen Meng. 2024. Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12577–12581.

[52] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4779–4783.

[53] Jiatong Shi, Yueqian Lin, Xinyi Bai, Keyi Zhang, Yuning Wu, Yuxun Tang, Yifeng Yu, Qin Jin, and Shinji Watanabe. 2024. Singing Voice Data Scaling-up: An Introduction to ACE-Opencpop and ACE-KiSing. In *Proc. Interspeech 2024*. 1880–1884.

[54] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567* (2020).

[55] svc-develop-team. 2023. so-vits-svc. https://github.com/svc-develop-team/so-vits-svc.

[56] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA.

[57] Xin Wang, Hector Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. 2024. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739* (2024).

[58] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. 2022. Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. In *Proc. Interspeech 2022*. 4242–4246.

[59] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xiaopeng Wang, Yuankun Xie, Xin Qi, Shuchen Shi, Yi Lu, Yukun Liu, et al. 2025. Mixture of experts fusion for fake audio detection using frozen wav2vec 2.0. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[60] Zhaolin Wei, Dengpan Ye, Jiacheng Deng, and Yuhan Lin. 2025. From Voices to Beats: Enhancing Music Deepfake Detection by Identifying Forgeries in Background. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[61] Yuankun Xie, Ruibo Fu, Zhiyong Wang, Xiaopeng Wang, Songjun Cao, Long Ma, Haonan Cheng, and Long Ye. 2026. Detect All-Type Deepfake Audio: Wavelet Prompt Tuning for Enhanced Auditory Perception. *Proceedings of the AAAI Conference on Artificial Intelligence* (2026).

[62] Yuankun Xie, Xiaoxuan Guo, Jiayi Zhou, Tao Wang, Jian Liu, Ruibo Fu, Xiaopeng Wang, Haonan Cheng, and Long Ye. 2026. Interpretable All-Type Audio Deepfake Detection with Audio LLMs via Frequency-Time Reinforcement Learning. *arXiv preprint arXiv:2601.02983* (2026).

[63] Yuxiong Xu, Bin Li, Weixiang Li, Sara Mandelli, Viola Negroni, and Sheng Li. 2025. ALDEN: Dual-Level Disentanglement with Meta-learning for Generalizable Audio Deepfake Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 7277–7286.

[64] Jun Xue, Yi Chai, Yanzhen Ren, Jinshen He, Zhiqiang Tang, Zhuolin Yi, Yihuan Huang, Yuankun Xie, and Yujie Chen. 2026. Unifying Speech Editing Detection and Content Localization via Prior-Enhanced Audio LLMs. *arXiv preprint arXiv:2601.21463* (2026).

[65] Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, et al. 2025. LLaSA: Scaling Train-Time and Inference-Time Compute for LLaMA-Based Speech Synthesis. *arXiv preprint arXiv:2502.04128* (2025).

[66] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *Proceedings of ICASSP*. IEEE, 9216–9220.

[67] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chuyuan Zhang, Xiaohui Zhang, Zhao Yan, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, and Haizhou Li. 2023. ADD 2023: the Second Audio Deepfake Detection Challenge. *ADD 2023: the Second Audio Deepfake Detection Challenge, accepted by IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis* (2023).

[68] Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, Haohe Liu, Wenwu Wang, and Mark D Plumbley. 2025. EnvSDD: Benchmarking Environmental Sound Deepfake Detection. In *Interspeech 2025*. 201–205. doi:10.21437/Interspeech.2025-1143

[69] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. 2022. M4Singer: a multi-style, multi-singer and musical score provided mandarin singing corpus. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 6914–6926.

[70] Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio deepfake detection with self-supervised XLS-R and SLS classifier. In *ACM Multimedia 2024*.

[71] Qishan Zhang, Shuangbing Wen, Fangke Yan, Tao Hu, and Jun Li. 2024. XWSB: A Blend System Utilizing XLS-R and Wavlm With SLS Classifier Detection System for SVDD 2024 Challenge. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 788–794.

[72] Tong Zhang, Yihuan Huang, and Yanzhen Ren. 2025. EchoFake: A Replay-Aware Dataset for Practical Speech Deepfake Detection. *arXiv preprint arXiv:2510.19414* (2025).

[73] Xueping Zhang, Han Yin, Yang Xiao, Lin Zhang, Ting Dang, Rohan Kumar Das, and Ming Li. 2026. ESDD2: Environment-Aware Speech and Sound Deepfake Detection Challenge Evaluation Plan. *arXiv preprint arXiv:2601.07303* (2026).

[74] You Zhang, Yongyi Zang, Jiatong Shi, Ryuichi Yamamoto, Tomoki Toda, and Zhiyao Duan. 2024. Svdd 2024: The inaugural singing voice deepfake detection challenge. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 782–787.

[75] Zirui Zhang, Wei Hao, Aroon Sankoh, William Lin, Emanuel Mendiola-Ortiz, Junfeng Yang, and Chengzhi Mao. 2025. I Can Hear You: Selective Robust Training for Deepfake Audio Detection. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=2GcR9bO620

[76] Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. MeloTTS: High-Quality Multi-Lingual Multi-Accent Text-to-Speech. https://github.com/myshell-ai/MeloTTS. GitHub repository.