A Neural Codec-Aware Training Strategy for Robust Audio Deepfake Detection

GPT-5.2, Claude Sonnet 4.5, Claude Opus 4.5 and Liner

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural audio codecs are increasingly adopted in modern speech processing systems; however, codec-induced artifacts can severely undermine the reliability of audio deepfake detection. In particular, bona fide speech processed by neural codecs is frequently misclassified as spoof, leading to critical false rejections in practical scenarios. Despite this emerging challenge, the treatment and labeling of neural-codec-processed speech in deepfake detection remain ambiguous. In this work, we systematically analyze the impact of neural audio codec processing on audio deepfake detection and clarify labeling criteria based on a content-origin definition of bona fide speech. Building on this formulation, we propose a codec-aware training strategy that explicitly models the relationship among clean bona fide speech, codec-processed bona fide speech, and spoof samples. Experimental results across multiple datasets and state-of-the-art detection models demonstrate that the proposed approach substantially improves robustness to neural-codec-processed speech, while preserving reliable performance on clean audio. These findings highlight the importance of codec-aware training for robust and realistic audio deepfake detection.

1 Introduction

Recent advances in artificial intelligence have made it significantly easier to generate high-quality audio deepfakes than in the past Wu et al. [2015]. In particular, deep learning–based text-to-speech (TTS) Zhang et al. [2022], Ju et al. [2024] synthesis and voice conversion (VC) Choi et al. [2024], Yao et al. [2025] technologies leverage large-scale speech datasets and powerful neural network models to accurately replicate fine-grained speaker characteristics such as pronunciation, prosody, emotional expression, and timbre. These developments have greatly simplified and automated the voice forgery process, which previously required substantial expertise and computational resources, enabling even non-expert users to produce highly realistic deepfake audio at low cost and within a short time. As a result, audio deepfakes now pose a growing challenge, as they are increasingly difficult to distinguish from genuine speech and raise serious concerns related to voice fraud, misinformation, and identity abuse, alongside their potential beneficial applications Hery et al. [2024].

Because of the growing risks associated with audio deepfakes, detection technologies have received increasing attention from both the research community and industry. As synthetic speech becomes more realistic and accessible, reliable audio deepfake detection is essential to safeguard biometric authentication systems, prevent voice-based fraud, and maintain trust in spoken media. To accelerate progress in this area, large-scale benchmarking initiatives such as the ASVspoof Challenge Todisco

Preprint. Under review.

et al. [2019], Yamagishi et al. [2021] and the Audio Deepfake Detection Challenge Yi et al. [2022, 2023a] have been organized. These challenges provide standardized datasets, evaluation protocols, and competitive settings that encourage the development and fair comparison of robust detection algorithms, thereby playing a crucial role in advancing the state of the art in audio deepfake detection Tak et al. [2022], Rosello et al. [2023], Truong et al. [2024].

In parallel with advances in speech synthesis and detection, recent progress in deep learning has also driven rapid development of neural audio codec technologies Défossez et al. [2022]. Unlike traditional signal-processing–based codecs, neural audio codecs compress and reconstruct speech using neural networks, enabling high perceptual quality at very low bitrates. As a result, they are increasingly integrated into modern speech generation, transmission, and synthesis pipelines Xin et al. [2024a,b].

However, this widespread adoption of neural audio codecs introduces new challenges for audio deepfake detection Yi et al. [2023b]. Many existing detection approaches rely on identifying subtle artifacts left by neural networks during the generation of synthetic speech Tak et al. [2021], Jung et al. [2022]. When bona fide speech produced by real speakers undergoes neural audio codec processing, codec-induced artifacts can closely resemble those introduced by generative models Lu et al. [2024]. Consequently, codec-processed bona fide speech is often misclassified as spoofed, leading to a substantial increase in false positives and degraded detection performance Zhang et al. [2025]. These effects blur the representational boundary between bona fide and spoofed speech, revealing a fundamental limitation of existing detection systems under neural codec environments.

Beyond performance degradation, neural audio codec processing also raises an important yet insufficiently addressed question regarding the labeling of codec-processed speech in audio deepfake detection. While some recent studies implicitly treat any speech that has passed through neural network–based processing as fake, such an assumption becomes increasingly restrictive as neural codecs are deployed in real-world communication systems. From a practical perspective, speech that originates from a real human speaker should remain bona fide even if it undergoes neural audio codec processing, provided that the underlying content is preserved. This ambiguity in labeling highlights the need for detection frameworks that explicitly distinguish content origin from signal processing artifacts.

In this paper, we first systematically demonstrate that neural audio codec processing significantly degrades the ability of existing audio deepfake detection systems to correctly identify bona fide speech. We then clarify labeling criteria for codec-processed speech by adopting a content-origin–based definition of bona fide speech. Based on this formulation, we propose a codec-aware training strategy that explicitly models the relationship among clean bona fide speech, neural-codec-processed bona fide speech, and spoof samples. Extensive experiments across multiple datasets and state-of-the-art detection models show that the proposed approach substantially improves robustness to neural codec processing while maintaining reliable performance on clean speech.

2 Related work

2.1 Neural audio codecs and speech processing

Neural audio codecs have recently emerged as a powerful alternative to traditional signal-processing–based speech codecs, leveraging deep neural networks to achieve high perceptual quality at extremely low bitrates Défossez et al. [2022]. By learning compact latent representations of speech Xin et al. [2024b], neural codecs enable efficient compression and high-fidelity reconstruction, and are increasingly adopted in speech synthesis, voice conversion, and speech communication systems Zhang et al. [2022]. These codecs are often integrated into end-to-end neural pipelines, making them a key component in modern audio generation and transmission frameworks Zeghidour et al. [2021]. While prior studies have primarily focused on improving reconstruction quality and compression efficiency, the implications of neural audio codecs for downstream tasks—particularly from a security and forensic perspective—remain relatively underexplored. In particular, their impact on audio deepfake detection performance has not been sufficiently investigated, despite the growing prevalence of neural codecs in real-world speech processing pipelines.

2.2 Audio deepfake detection

Audio deepfake detection has attracted significant attention due to rapid advances in speech synthesis and voice conversion technologies. Early studies relied on hand-crafted acoustic features combined with conventional classifiers to capture artifacts introduced during speech generation Wu et al. [2014a], Sahidullah et al. [2015]. More recent approaches predominantly employ deep learning–based models, such as convolutional neural networks and transformer architectures Baevski et al. [2020], Chen et al. [2022], to automatically learn discriminative representations from raw waveforms or time–frequency features. Large-scale evaluation campaigns Yamagishi et al. [2021], Yi et al. [2023a] have further accelerated progress by providing standardized datasets and benchmarking protocols. Despite these advances, most existing detection methods are developed and evaluated under relatively controlled conditions, often assuming clean signals or conventional signal distortions. As a result, their robustness to emerging speech processing techniques, including neural audio codecs, remains limited, leading to significant performance degradation in more realistic deployment scenarios.

2.3 Training Objectives for Audio Deepfake Detection

Most audio deepfake detection systems formulate the task as a binary classification problem and are trained using standard cross-entropy–based objectives. While effective for optimizing decision-level discrimination between genuine and synthetic speech, such training objectives primarily emphasize label supervision and may fail to encourage robust and invariant feature representations. This limitation becomes particularly evident under distribution shifts caused by signal processing operations, such as compression or codec transformations. Recent studies suggest that improving robustness requires training strategies that go beyond conventional classification losses by incorporating additional constraints or complementary supervisory signals Doan et al. [2024b,a]. However, existing approaches do not explicitly account for distortions introduced by neural audio codecs, motivating further exploration of codec-aware training objectives tailored for robust audio deepfake detection.

3 Motivation

In this section, we empirically investigate whether neural audio codec processing indeed degrades the performance of existing audio deepfake detection systems, as discussed in the Introduction and Related Work. Our goal is not to propose a solution at this stage, but to provide experimental evidence that neural codec–processed bona fide speech poses a significant challenge to current detectors.

3.1 Experimental Setup for Motivation Analysis

To obtain representative observations, we conduct this analysis using three state-of-the-art audio deepfake detection models that are widely adopted as benchmarks in the literature: SSL-AASIST, SSL-Conformer, and SSL-Conformer-TCM. We evaluate these models on six datasets covering diverse recording conditions and attack scenarios. Bona fide speech is drawn from VCTK, LibriSpeech, and VoxCeleb, while spoofed speech is collected from ASVspoof 2021 DF, DSD-Corpus, and an In-the-Wild dataset. Detailed descriptions of the models and datasets are provided in section **??**.

To isolate the impact of neural audio codec processing, we apply two recent neural audio codecs, BigCodec and SpeechTokenizer, to bona fide samples only, while keeping spoof samples unchanged. From each dataset, 5,000 utterances are randomly sampled, resulting in a total of 30,000 evaluation samples. Starting from a setting with no codec-processed bona fide speech, we progressively replace a portion of the bona fide samples with their neural-codec-processed counterparts, increasing the replacement ratio (e.g., 20%, 40%, and beyond) while maintaining a fixed number of spoof samples. This evaluation protocol allows us to systematically examine how detection performance changes as neural codec processing becomes increasingly prevalent, thereby providing direct empirical evidence of its impact on existing detection systems.

Evaluation Metrics. We evaluate audio deepfake detection performance using Accuracy (ACC) and Equal Error Rate (EER), which are standard metrics in spoofing and deepfake detection tasks.

ACC is defined as

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

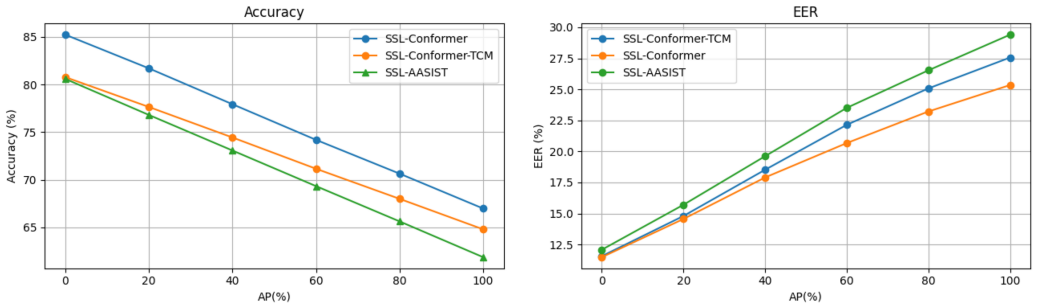

Figure 1: Detection performance trends as the proportion of neural-codec-processed bona fide samples increases. ACC consistently decreases while EER increases across all models, indicating that neural codec processing of bona fide speech leads to severe performance degradation. The x-axis value of 0% corresponds to evaluation using only original (non–codec-processed) data. At each inclusion ratio, an equal number of samples processed by different neural codecs are incorporated.

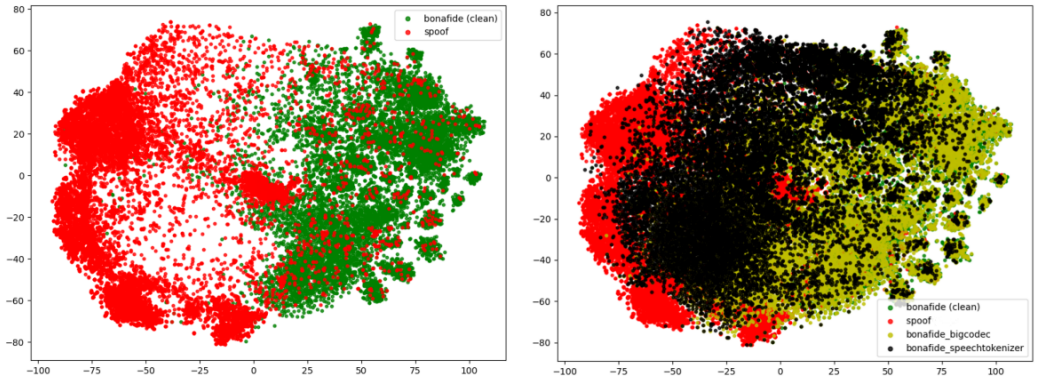

Figure 2: t-SNE visualization of embedding distributions without (left) and with (right) neural audio codec–processed bona fide samples. When neural codec processing is introduced, bona fide embeddings substantially overlap with spoof samples, indicating codec-induced representation drift that degrades detection performance.

where TP and TN denote the numbers of correctly classified bona fide and spoof samples, respectively, and FP and FN represent false positives and false negatives.

The EER is defined as the operating point at which the false acceptance rate (FAR) equals the false rejection rate (FRR):

$$\text{EER} = \text{FAR}(\tau^*) = \text{FRR}(\tau^*), \tag{2}$$

where τ^* denotes the decision threshold.

3.2 Impact of Neural Audio Codecs on Audio Deepfake Detection

Figure 1 illustrates the impact of neural audio codec sample inclusion on audio deepfake detection performance in terms of ACC and EER. Across all evaluated models, a consistent trend is observed: as the proportion of neural-codec-processed bona fide samples increases, detection accuracy degrades substantially, while EER increases sharply.

This behavior indicates that codec-processed bona fide speech is increasingly misclassified as spoofed, leading to a substantial rise in false positives. Notably, this degradation is consistent across different model architectures, suggesting that the observed performance drop is not model-specific but rather reflects a fundamental limitation of existing detection approaches when exposed to neural audio codec processing. These results provide direct empirical evidence that applying neural audio codecs to bona fide speech poses a serious challenge to current audio deepfake detection systems.

To further analyze the underlying cause of the performance degradation observed in Figure 1, we visualize the embedding space using t-SNE Maaten and Hinton [2008]. Figure 2 compares the embedding distributions obtained without neural audio codec processing (left) and with neural-codec-processed bona fide samples included (right).

In the absence of neural audio codec processing, bona fide and spoof samples form relatively well-separated clusters in the embedding space. In contrast, when neural-codec-processed bona fide samples are introduced, their embeddings substantially overlap with those of spoof samples. This overlap indicates that codec-induced distortions shift bona fide representations toward spoof regions, thereby blurring the decision boundary learned by the detector.

Taken together, these observations demonstrate that neural audio codec processing disrupts the representation space in a manner that directly contributes to increased false positives and degraded detection performance. This representation-level evidence complements the trends observed in decision-level metrics and highlights the need for training strategies that explicitly preserve the proximity between clean and codec-processed bona fide speech while maintaining a clear separation from spoofed speech, which we address in the following section.

4 Method

In this section, we present a training strategy for improving the robustness of audio deepfake detection under neural audio codec processing. We first introduce the problem formulation and labeling strategy adopted in this study, followed by an overview of the proposed training framework. Our approach is motivated by the limitation of standard cross-entropy–based classification, which often fails to correctly handle bona fide speech that has undergone neural audio codec processing.

To mitigate this issue, we employ two complementary loss functions that guide the model to better distinguish codec-processed bona fide speech from spoofed speech. A modified softplus loss Dugas et al. [2000] encourages codec-processed samples to remain closer to their corresponding original bona fide samples than to spoof samples, thereby reducing false positives at the decision level. In addition, a triplet loss Schroff et al. [2015] is used to explicitly structure the embedding space by pulling codec-processed samples toward their original bona fide counterparts while pushing spoof samples farther away, with respect to the original bona fide speech. The detailed formulations and optimization procedures of these loss functions are described in the subsequent subsections.

4.1 Problem formulation and label definition

Problem formulation. Under neural audio codec processing, audio deepfake detection is formulated as a binary classification problem that aims to distinguish between genuine human speech (bona fide) and artificially generated or manipulated speech (spoof). Given an input audio signal

$$x \in \mathbb{R}^T, \tag{3}$$

where T denotes the temporal length of the waveform, the objective is to learn a detection function

$$f : \mathbb{R}^T \to \{0, 1\}, \tag{4}$$

which predicts a binary label y indicating whether the input signal is bona fide or spoof:

$$y = f(x), \tag{5}$$

where

$$y = \begin{cases} 1, & \text{if } x \text{ is bona fide (real)}, \\ 0, & \text{if } x \text{ is spoof (fake)}. \end{cases} \tag{6}$$

The detection model is trained in a supervised manner using labeled datasets containing both bona fide and spoofed speech samples.

Label definition. In the ASVspoof community, the labeling strategy is primarily based on the origin of speech content. Speech is labeled as bona fide (real) if it is produced by a genuine human speaker through a natural speaking process Delgado et al. [2021], whereas speech is labeled as spoof (fake) if it is generated or altered by artificial systems such as text-to-speech, voice conversion, or replay attacks Wu et al. [2014b]. Under this definition, post-processing operations that do not alter the

semantic content or speaker origin of the speech—such as channel effects or conventional codecs—do not change the original class label.

In contrast, recent studies such as Codecfake Lu et al. [2024] adopt a different perspective, where any audio signal that has passed through a neural network–based processing module is regarded as fake, regardless of whether the underlying speech content originates from a real human speaker. From this viewpoint, neural audio codecs are treated as generative models that introduce artificial artifacts, and codec-processed speech is therefore labeled as spoof. While this definition is useful for analyzing vulnerabilities specific to neural processing pipelines, it implicitly assumes that the presence of a neural network alone is sufficient to redefine speech authenticity.

In this work, we argue that such an assumption becomes increasingly restrictive as neural networks are progressively integrated into real-world speech communication systems. As neural audio codecs and other neural speech processing modules are expected to be widely deployed in practical applications, treating all neural-network-processed speech as fake does not align with realistic usage scenarios. Instead, we follow the ASVspoof community's content-origin–based definition and consider the origin of the speech content as the primary criterion for labeling.

Accordingly, in our study, speech remains labeled as bona fide as long as it originates from a real human speaker, even if it undergoes neural audio codec processing, provided that the semantic content and speaker identity are preserved. To explicitly account for codec processing while maintaining label consistency, we introduce an auxiliary label, denoted as NC_bona, to represent bona fide speech that has passed through a neural codec. Importantly, this auxiliary label is used only to define relative roles in the training objective and for analysis purposes, while the original class label remains bona fide during supervision. This labeling strategy enables systematic investigation of neural codec effects while remaining consistent with realistic deployment conditions and established evaluation protocols.

4.2 Overall training framework

Conventional audio deepfake detection models are typically trained using binary supervision, where each utterance is labeled as either bona fide or spoof. Such a formulation does not explicitly account for neural audio codec processing, which can significantly distort the feature space and cause codec-processed bona fide speech to overlap with spoof samples, as demonstrated in Section 3.

To address this limitation, we propose a codec-aware training framework that explicitly models the relationship among clean bona fide speech, neural-codec-processed bona fide speech (denoted as NC_bona), and spoofed speech. Let x^b denote a clean bona fide utterance produced by a real speaker, x^{nc} denote its corresponding neural-codec-processed counterpart, and x^s denote a spoofed utterance. All samples are passed through a shared embedding network $f(\cdot)$ to obtain

$$z = f(x), \quad z \in \mathbb{R}^D, \tag{7}$$

where D denotes the embedding dimension.

During training, mini-batches are constructed using an anchor–positive–negative scheme. Specifically, each mini-batch contains a set of clean bona fide anchors $\{x^b\}$, multiple NC_bona positives $\{x^{nc}\}$ corresponding to each anchor, and spoofed negatives $\{x^s\}$. Although clean bona fide and NC_bona samples share the same ground-truth class label for classification, they are assigned distinct roles in the training objective to explicitly encourage codec-robust learning.

The proposed framework optimizes two complementary objectives that address different failure modes identified in Section 3. First, a modified softplus-based separation loss operates at the decision level, encouraging codec-processed bona fide speech to be classified closer to its original bona fide counterpart than to spoofed speech, thereby directly reducing false positives caused by codec-induced artifacts. Second, a triplet loss operates at the representation level, explicitly structuring the embedding space by pulling codec-processed samples toward their corresponding clean bona fide anchors while pushing spoof samples farther away. Together, these objectives jointly preserve intra-class consistency between clean and codec-processed bona fide speech while maintaining a clear separation from spoofed speech. The detailed formulations of these losses are described in the following subsections.

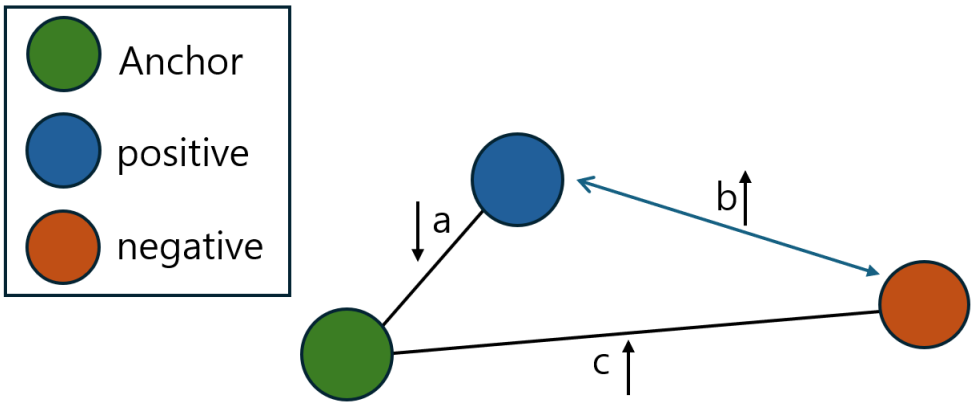

Figure 3: Illustration of how the separation loss and the triplet loss jointly shape the embedding space. The separation loss pulls positive samples toward the anchor while pushing them away from negative samples, and the triplet loss enforces a margin between the anchor and the negative sample.

4.3 Modified softplus loss for deepfake detection

The first objective of our training framework is to explicitly separate neural-codec-processed bona fide speech from spoofed speech in the embedding space. As shown in Section 3, codec-processed bona fide samples often lie close to spoof samples, making them particularly difficult to classify using conventional classification losses.

To address this challenge, we introduce a softplus-based separation loss that focuses on hard positive samples. For a given clean bona fide anchor x^b, we define the set of positive samples $\mathcal{P}(x^b)$ as the codec-processed bona fide samples obtained from different neural codecs within the same mini-batch. Among these, the hard positive x^{hp} is defined as the positive sample farthest from the anchor in the embedding space:

$$x^{hp} = \arg \max_{p \in \mathcal{P}(x^b)} d\big(f(x^b), f(p)\big), \tag{8}$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance.

Let x^s denote a spoofed negative sample. To enforce an explicit separation margin $m > 0$, we impose the following constraint:

$$d\big(f(x^{hp}), f(x^b)\big) + m < d\big(f(x^{hp}), f(x^s)\big). \tag{9}$$

This constraint encourages codec-processed bona fide samples to remain closer to their clean bona fide anchors than to spoof samples by at least a margin m. As illustrated in Figure 3, the separation loss compares two distances: $a = d(f(x^{hp}), f(x^b))$ between the clean anchor and the hard positive, and $b = d(f(x^{hp}), f(x^s))$ between the hard positive and a spoofed sample. Enforcing $a + m < b$ (equivalently, $b - a \geq m$) drives the codec-processed bona fide sample to stay close to the anchor while being sufficiently separated from spoof samples.

This constraint is optimized using the following softplus loss:

$$\mathcal{L}_{\text{sep}} = \log\big(1 + \exp(\Delta)\big), \quad \Delta = d\big(f(x^{hp}), f(x^b)\big) - d\big(f(x^{hp}), f(x^s)\big) + m. \tag{10}$$

By leveraging the smooth gradient properties of the softplus function, this loss emphasizes difficult codec-processed samples while enforcing a controlled separation margin, thereby avoiding overly aggressive updates and improving robustness to neural codec–induced distortions.

4.4 Triplet loss

While the softplus-based separation loss explicitly focuses on separating codec-processed bona fide samples from spoofed speech, it does not directly constrain the anchor-centered geometry between clean bona fide speech and spoofed speech in the embedding space. In particular, separation loss

alone may be insufficient to preserve stable anchor-level decision boundaries when codec-induced distortions are severe. To address this limitation, we incorporate a triplet loss that enforces a margin-based separation centered at the clean bona fide anchor.

Each triplet consists of a clean bona fide anchor x^b, a positive sample x^{nc} (neural-codec-processed bona fide), and a negative sample x^s (spoof). The triplet loss is defined as

$$\mathcal{L}_{\text{tri}} = \Big[\|f(x^b) - f(x^{nc})\|_2^2 - \|f(x^b) - f(x^s)\|_2^2 + \alpha\Big]_+, \tag{11}$$

where $\alpha > 0$ is a margin hyperparameter and $[\cdot]_+$ denotes the hinge function.

As illustrated in Figure 3, the triplet loss compares the anchor-centered distances $a = d\big(f(x^b), f(x^{nc})\big)$ (anchor–positive) and $c = d\big(f(x^b), f(x^s)\big)$ (anchor–negative). When the constraint $a + \alpha < c$ (equivalently, $c - a \geq \alpha$) is violated, the hinge term becomes active, driving the model to reduce a while increasing c.

This objective encourages codec-processed bona fide samples to remain close to their clean anchors while pushing spoof samples farther away by at least the margin α, thereby stabilizing anchor-centered decision boundaries under neural audio codec processing. In contrast to the codec-focused separation loss, the triplet loss explicitly regularizes the anchor-centered geometry of the embedding space and helps prevent degenerate or drifting representations.

4.5 Joint optimization strategy

The final training objective is designed to jointly optimize decision-level discrimination and codec-robust representation learning. To this end, we combine the softplus-based separation loss, the triplet loss, and a conventional classification objective.

Specifically, the separation loss $\mathcal{L}_{\text{sep}}$ focuses on resolving hard confusion cases by pushing neural-codec-processed bona fide samples away from spoof samples while maintaining proximity to their corresponding clean anchors. In parallel, the triplet loss $\mathcal{L}_{\text{tri}}$ preserves anchor-centered decision boundaries by explicitly regularizing the relative geometry between clean bona fide speech, codec-processed counterparts, and spoofed speech in the embedding space. These two objectives address complementary failure modes observed under neural codec processing, as demonstrated in Section 3.

In addition to these objectives, we incorporate a standard cross-entropy loss $\mathcal{L}_{\text{ce}}$, which is widely used in audio deepfake detection, to provide stable decision-level supervision between bona fide and spoof classes. The cross-entropy loss is defined as

$$\mathcal{L}_{\text{ce}} = -\log \frac{\exp(s_y)}{\sum_{k\in\{0,1\}} \exp(s_k)}, \tag{12}$$

where $s = h(f(x)) \in \mathbb{R}^2$ denotes the logits produced by the classification head $h(\cdot)$, and $y \in \{0, 1\}$ is the ground-truth label.

The overall training objective is formulated as

$$\mathcal{L} = \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{ce}}, \tag{13}$$

where λ_{sep} and λ_{tri} control the relative contributions of the separation loss and the triplet loss, respectively.

By jointly optimizing these three complementary objectives, the proposed framework simultaneously achieves codec-focused separation, anchor-level boundary preservation, and stable class discrimination. This joint optimization strategy enables robust audio deepfake detection under neural audio codec processing while maintaining compatibility with standard evaluation protocols.

5 Evaluation

In this section, we evaluate the effectiveness of the proposed codec-aware training framework for audio deepfake detection under neural audio codec processing.

5.1 Experimental setup

Training data. For model training, we use the ASVspoof 2019 Logical Access (LA19) dataset Nautsch et al. [2021], including both the training and development splits. To improve robustness against neural audio codec processing while maintaining label consistency as discussed in Section 4.1, we augment the training data by applying neural audio codecs only to bona fide samples. Specifically, we employ two representative neural codecs, BigCodec Xin et al. [2024a] and SpeechTokenizer Xin et al. [2024b], to generate codec-processed bona fide speech while preserving the original ground-truth labels.

Evaluation data. To evaluate generalization performance under diverse and realistic conditions, we construct an evaluation set using multiple datasets for both bona fide and spoof classes. For bona fide speech, we use samples from LibriSpeech **?**, VCTK Veaux et al. [2017], and VoxCeleb Nagrani et al. [2017], which provide diverse speaker characteristics and recording conditions. For spoofed speech, we use three datasets: ASVspoof 2021 DF eval Liu et al. [2023], which serves as an official benchmark; DSD-Corpus **?**, which includes a wide range of synthetic speech generated by different synthesis and voice conversion systems; and the In-the-Wild dataset **?**, consisting of real-world audio collected from social media platforms.

From each dataset, we randomly select 5,000 samples. In addition to the original signals, we apply neural audio codecs to all selected samples, resulting in both clean and codec-processed versions. This evaluation protocol yields a total of 90,000 evaluation samples and allows us to assess detection robustness under both clean and codec-processed conditions, including challenging scenarios where codec effects are present during deployment.

Detection models. We evaluate the proposed training strategy using several state-of-the-art audio deepfake detection models. Specifically, we employ SSL-AASIST Tak et al. [2022], SSL-Conformer Rosello et al. [2023], and SSL-Conformer-TCM Truong et al. [2024] as backbone architectures. All detection models share a common front-end based on self-supervised learning (SSL). We use a wav2vec 2.0–based XLSR model **??** for feature extraction, which has demonstrated strong performance in audio deepfake detection. The extracted frame-level representations are then fed into different backend architectures. SSL-AASIST utilizes a graph neural network (GNN)–based classifier to capture relational structures, while SSL-Conformer combines convolutional and transformer modules to model both local and global temporal dependencies. SSL-Conformer-TCM further incorporates temporal-channel modeling to enhance robustness against temporal distortions.

Hyperparameter settings. For the metric-learning objectives, we sweep margin hyperparameters. For the triplet loss, we set the margin to $\alpha \in \{0.2, 0.3, 0.5\}$, following Schroff et al. [2015], with $\alpha = 0.2$ as the default value. For the separation loss, we sweep the margin m and use $m = 0.5$ as the default. For the joint objective in Eq. (13), we fix the loss weights to $\lambda_{\text{sep}} = \lambda_{\text{tri}} = 1$ for simplicity, and analyze sensitivity to these hyperparameters in the ablation study.

5.2 Results

Table 1 summarizes codec-wise accuracy under Original, BigCodec, and SpeechTokenizer conditions for the baseline, simple augmentation, and the proposed training strategy. Under the baseline setting, applying neural audio codec processing leads to a clear and consistent performance degradation across all backbones, indicating that models trained solely on clean data are highly sensitive to codec-induced distortions.

When simple data augmentation with neural-codec-processed bona fide samples is applied, performance under codec conditions is partially improved. However, this improvement comes at the cost of degraded performance on Original audio across all backbones, demonstrating that naive augmentation alone fails to achieve a balanced optimization between clean and codec-processed domains and may disrupt decision boundaries learned from clean data.

In contrast, training with the proposed codec-aware strategy consistently improves performance across nearly all conditions, including Original, BigCodec, and SpeechTokenizer, for all evaluated backbones. Importantly, these gains are achieved without sacrificing performance on clean audio, resulting in substantial improvements in average accuracy. For example, under the SpeechTokenizer

condition, the proposed method achieves absolute accuracy gains of up to 19% compared to the baseline, highlighting its effectiveness in addressing severe codec-induced distortions.

Notably, the most significant improvements are observed under the SpeechTokenizer condition, which represents the most challenging scenario for baseline models. As illustrated in Figure 2 (right), SpeechTokenizer-processed bona fide samples tend to lie closer to spoof samples in the embedding space, causing them to be frequently selected as hard positives during training. Consequently, the proposed separation and triplet objectives place greater emphasis on these difficult cases, leading to more focused and effective optimization.

Overall, the consistent performance gains across all backbones and codec conditions, as well as the improved average accuracy, demonstrate the effectiveness of the proposed training strategy in mitigating neural-codec-induced performance degradation while preserving reliable detection performance on clean audio.

Table 1: Accuracy (%) across codec processing conditions for the baseline and the proposed method

(a) SSL-Conformer

Margins		Accuracy (%)			
Triplet(α)	Sep.(m)	Original	BigCodec	SpeechTokenizer	Avg.
Baseline		85.43	77.69	65.29	76.14
Augmentation		82.02	75.23	74.00	77.08
0.2	0.5	86.22	77.41	80.53	81.39
0.3	0.5	90.32	82.49	84.61	85.81
0.5	0.5	89.84	80.55	82.20	84.20

(b) SSL-Conformer-TCM

Margins		Accuracy (%)			
Triplet(α)	Sep.(m)	Original	BigCodec	SpeechTokenizer	Avg.
Baseline		80.77	74.85	62.25	72.62
Augmentation		80.29	73.06	72.22	75.19
0.2	0.5	86.96	77.35	80.48	81.60
0.3	0.5	82.91	74.63	77.26	78.27
0.5	0.5	85.40	75.94	79.64	80.33

(c) SSL-AASIST

Margins		Accuracy (%)			
Triplet(α)	Sep.(m)	Original	BigCodec	SpeechTokenizer	Avg.
Baseline		80.58	73.19	55.86	69.88
Augmentation		73.40	68.41	65.14	68.98
0.2	0.5	88.04	78.25	78.20	81.50
0.3	0.5	87.85	76.68	78.22	80.92
0.5	0.5	85.30	72.95	74.45	77.57

5.3 Bona fide-side analysis under neural codec processing

To specifically analyze bona fide-side behavior under neural audio codec processing, Table 2 reports accuracy evaluated only on bona fide speech, separately for clean (Original) bona fide samples and neural-codec-processed bona fide samples (NC_bona). This analysis directly reveals how different

Table 2: Bona fide accuracy (%) on clean and neural codec processed bona fide speech (NC_bona)

Margins		SSL-Conformer		SSL-AASIST		SSL-Conformer-TCM	
Triplet(α)	Sep.(m)	Bona fide	NC_bona	Bona fide	NC_bona	Bona fide	NC_bona
Baseline		77.10	49.27	65.84	32.93	66.36	41.42
Augmentation		73.14	57.03	51.16	37.22	67.14	51.84
0.2	0.5	89.16	74.09	92.36	70.78	91.92	75.64
0.3	0.5	96.39	81.43	88.55	67.60	84.88	74.64
0.5	0.5	91.88	73.71	82.58	58.85	92.57	78.46

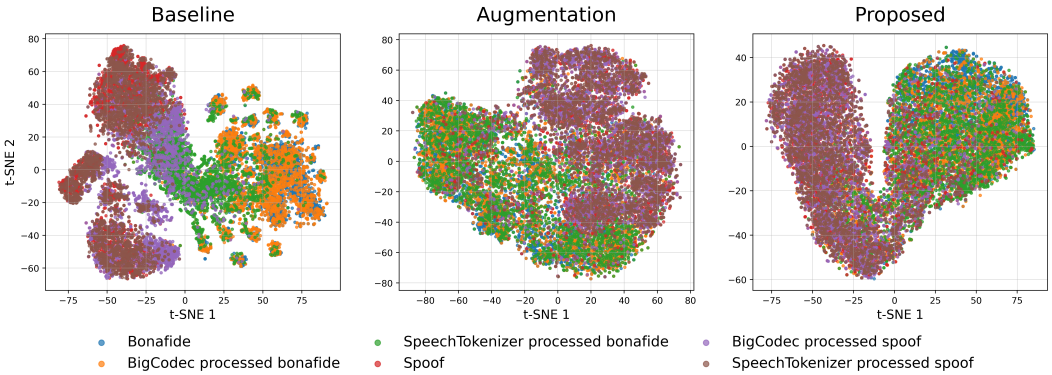

Figure 4: t-SNE visualization of embedding distributions for the baseline, augmentation-based training, and the proposed training strategy. The proposed method yields clearer separation between bona fide and spoof samples, particularly under neural codec processing.

training strategies affect false positive behavior under codec-induced distortions, complementing the overall results presented in Section 5.2.

Under the baseline setting, bona fide accuracy drops drastically when neural codec processing is applied. In particular, accuracy on NC_bona decreases by more than 30–40 percentage points across backbones, indicating that baseline models frequently misclassify codec-processed bona fide speech as spoof. This confirms that neural audio codec processing is a primary source of false positives in existing detection systems.

When simple data augmentation is applied, accuracy on NC_bona improves compared to the baseline, demonstrating that exposure to codec-processed samples during training partially alleviates codec-related degradation. However, this improvement comes at the cost of reduced accuracy on clean bona fide speech, revealing a clear trade-off between clean and codec-processed conditions. This suggests that naive augmentation fails to preserve consistent decision boundaries across domains.

In contrast, the proposed training strategy consistently improves bona fide accuracy for both Original and NC_bona samples across all backbones. Notably, the proposed method recovers a substantial portion of the degraded NC_bona performance, achieving absolute gains of up to 38 percentage points over the baseline, while simultaneously improving or maintaining accuracy on clean bona fide speech. These results demonstrate that explicitly modeling the relationship among clean bona fide, codec-processed bona fide, and spoof samples enables the model to learn codec-robust representations without sacrificing reliability on clean speech.

5.4 Analysis via embedding space

Figure 4 visualizes the embedding distributions using t-SNE for the baseline, augmentation-based training, and the proposed training strategy. Under the baseline setting, bona fide and spoof samples are heavily intermixed, particularly for neural-codec-processed bona fide samples. This observation

is consistent with the severe bona fide misclassification reported in Table 2, indicating that codec-induced distortions substantially blur the decision boundary in the embedding space.

Applying simple data augmentation slightly improves the separation between bona fide and spoof samples. However, substantial overlap remains, especially between codec-processed bona fide samples and spoof clusters. This suggests that naive augmentation provides limited structural separation and is insufficient to disentangle codec-induced variations at the representation level.

In contrast, the proposed training strategy yields a much clearer separation between bona fide and spoof regions. Both clean and codec-processed bona fide samples form more compact and coherent clusters that are well separated from spoof samples. This visualization suggests that the proposed codec-aware objectives effectively restructure the embedding space, reducing overlap caused by neural audio codec processing and leading to more discriminative and robust representations. These observations provide complementary representation-level evidence supporting the quantitative improvements reported in Sections 5.2 and 5.3.

6 Conclusion

This work investigated the degradation of audio deepfake detection performance under neural audio codec processing, with a particular focus on bona fide-side errors where codec-processed bona fide speech is misclassified as spoof. Our analysis demonstrated that neural audio codecs introduce distortions that fundamentally challenge existing detectors by blurring the representational boundary between bona fide and spoofed speech.

To address this issue, we proposed a codec-aware training strategy that explicitly models codec-induced variations during training and jointly optimizes complementary objectives at both the decision and representation levels. Extensive experiments across three representative backbone architectures showed that the proposed method consistently improves detection performance under neural codec processing, with especially strong gains under challenging conditions such as SpeechTokenizer.

Further analysis revealed that these improvements primarily stem from a substantial reduction in false positives on bona fide speech, enabling reliable bona fide recognition without sacrificing performance on clean audio. These findings highlight the importance of explicitly accounting for neural codec effects when designing robust and realistic audio deepfake detection systems.

Future work will extend this study to a broader range of neural codecs and system configurations, and explore generalization to unseen codec conditions encountered in real-world speech communication pipelines.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf`.

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

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17862–17870, 2024.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, et al. Asvspoof 2021: Automatic

speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535*, 2021.

Thien-Phuc Doan, Hung Dinh-Xuan, Taewon Ryu, Inho Kim, Woongjae Lee, Kihun Hong, and Souhwan Jung. Trident of poseidon: A generalized approach for detecting deepfake voices. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2222–2235, 2024a.

Thien-Phuc Doan, Long Nguyen-Vu, Kihun Hong, and Souhwan Jung. Balance, Multiple Augmentation, and Re-synthesis: A Triad Training Strategy for Enhanced Audio Deepfake Detection. In *Interspeech 2024*, pages 2105–2109, 2024b. doi: 10.21437/Interspeech.2024-24.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL `https://proceedings.neurips.cc/paper_files/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf`.

Andrew Hery, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. Audio deepfakes: Threats to voice assistants and voice-activated systems. 2024.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. Codecfake: An Initial Dataset for Detecting LLM-based Deepfake Audio. In *Interspeech 2024*, pages 1390–1394, 2024. doi: 10.21437/Interspeech.2024-2428.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

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

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. A Conformer-Based Classifier for Variable-Length Utterance Processing in Anti-Spoofing. In *Proc. INTERSPEECH 2023*, pages 5281–5285, 2023. doi: 10.21437/Interspeech.2023-1820.

Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. In *Interspeech 2015*, pages 2087–2091, 2015. doi: 10.21437/Interspeech.2015-472.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*, 2022.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541, 2024. doi: 10.21437/Interspeech.2024-659.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15, 2017.

Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–5. IEEE, 2014a.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, and Junichi Yamagishi. Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training*, 10(15): 3750, 2014b.

Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66: 130–153, 2015.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024a.

Z Xin, Z Dong, L Shimin, Z Yaqian, and Q Xipeng. Speechtokenizer: Unified speech tokenizer for speech language models. In *Proc. Int. Conf. Learn. Representations*, pages 1–21, 2024b.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

Jixun Yao, Yang Yuguang, Yu Pan, Ziqian Ning, Jianhao Ye, Hongbin Zhou, and Lei Xie. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25669–25677, 2025.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE, 2022.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*, 2023a.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023b.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Chengyi Zhang et al. Vall-e: Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2022.

You Zhang, Baotong Tian, Lin Zhang, and Zhiyao Duan. PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing . In *Interspeech 2025*, pages 5353–5357, 2025. doi: 10.21437/Interspeech.2025-942.