Adaptive Compression and Targeted Perturbation: A Unified Framework for Generalized Audio Deepfake Detection

Jianqiao Cui[†]
Yangtze Delta Region Institute of Tsinghua University
Jiaxing, China
jc18n17@soton.ac.uk

Haoxian Zhu
University of Southampton
Southampton, UK
h2v20@soton.ac.uk

Hanbiao Meng
Taiyuan University of Technology
Taiyuan, China
2024001543@link.tyut.edu.cn

Yanghao Zhang[†]
Imperial College London
London, UK
yanghao.zhang@imperial.ac.uk

Abstract

Existing audio deepfake detection (ADD) methods frequently exhibit limited generalization, as fixed bottleneck architectures fail to adapt to the heterogeneous information densities inherent in diverse spoofing artifacts. Moreover, standard augmentation often leads to a pathological robustness-accuracy trade-off. In this paper, we propose a unified ADD framework to address these challenges. Our approach integrates: (1) an Adaptive Audio Feature Integration (AAFI) module that dynamically selects optimal latent dimensions via a compression pool and a stochastic exploration mechanism, ensuring flexible representations across multiple types of dynamic attacks; and (2) a Mel-Band Adversarial Perturbation (Mel-BAP) strategy that applies targeted regularization in the mel-spectrogram domain to encourage the model to learn invariant decision boundaries. Extensive evaluations on benchmarks including ADD 2022, FakeOrReal, and In-the-Wild demonstrate that our model based on Whisper-small achieves a state-of-the-art average Equal Error Rate (EER) of 1.51%. Compared to existing SOTA models such as the 3-billion-parameter Resemble-Detect-3B-Omni, our framework reduces the parameter count by up to 91.6% while achieving a 60.7% improvement in average EER. Notably, the model exhibits exceptional zero-shot transferability, achieving EERs of 3.91–5.65% on unseen languages and maintaining strong resilience to real-world acoustic distortions. This work provides a scalable and efficient "plug-in" solution for enhancing trust and security in the audio Web ecosystem.

1 Introduction

Rapid advancements in speech generation [1, 2, 3] have enabled the integration of natural-sounding voices into virtual assistants and content creation [4, 5]. However, the misuse of these technologies through audio *deepfakes* poses severe security and privacy threats to the Web ecosystem, facilitating the spread of misinformation and causing reputational damage [6, 7, 8]. Consequently, robust detection methods are essential to maintain digital trust. While initiatives such as ASVspoof [9, 10] and ADD [11] have catalyzed algorithm development, cross-domain generalization remains a critical challenge. Detection models often suffer substantial performance degradation on unseen deepfakes

[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

[12] and across heterogeneous evaluation conditions [13]. Although models built on pre-trained speech representations, such as Wav2vec [14], have shown promising performance [14, 15, 16, 17], most apply the same fixed-dimensional transformation to every input, despite substantial variation in the salience and spectro-temporal complexity of spoofing artifacts. From the perspective of the trade-off between compression and fidelity associated with information bottleneck theory [18, 19], narrower representations may suppress task-irrelevant variation but risk discarding weak forensic cues, whereas wider representations may preserve subtle artifacts while retaining redundant or noisy information. A fixed representation width may therefore be suboptimal across the diverse noise levels, synthesis methods, and attack distributions encountered in practice [20, 21, 22, 23]. This motivates input-dependent multi-resolution representations that complement the original pre-trained features and adapt their compression level to heterogeneous spoofing conditions. Beyond unseen spoofing methods, adversarial perturbations pose an additional challenge to neural network robustness [24]. Prior work has explored adversarial training strategies that address training efficiency [25], adaptivity [26], and class-wise robustness disparities [27].

To address these challenges, we investigate the following research questions:

RQ1: Can input-dependent compressed representations complement frozen encoder features across heterogeneous spoofing conditions?
RQ2: How can we enhance robustness against unseen spoofing methods and adversarial attacks?

To investigate *RQ1*, we introduce the Adaptive Audio Feature Integration (AAFI) module. Rather than applying a fixed-width transformation to every input, AAFI uses an input-dependent router to select and fuse complementary representations from a heterogeneous compression pool. These multi-resolution features are combined with the original Whisper encoder representation to accommodate diverse spoofing conditions [28] while limiting information loss.

To enhance robustness (*RQ2*), we propose the *Mel-Band Adversarial Perturbation (Mel-BAP)* strategy. Recognizing that global augmentation can clutter the information bottleneck and sacrifice precision, we design Mel-BAP to perform adversarial training directly in the mel-spectrogram domain. By leveraging perceptually scaled frequency bands, it encourages the network to learn robust, perceptually aligned decision boundaries resilient to unseen spoofing patterns. We integrate AAFI and Mel-BAP into a unified framework and make the following contributions:

- **AAFI Module:** An adaptive integration mechanism with a compression pool to handle diverse information densities in dynamic attacks.
- **Mel-BAP Strategy:** A targeted frequency-domain adversarial training approach. Unlike generic frequency-selective methods, Mel-BAP is perceptually aligned with the Whisper backbone and exploits forensic priors associated with neural vocoders to achieve robustness.
- **State-of-the-art Performance:** Extensive tests on the ADD 2022, FakeOrReal, and In-the-Wild datasets show that our method outperforms current SOTA approaches by 28.8% with fewer trainable parameters.

2 Related Work

Audio Deepfake Datasets. We leverage two large-scale datasets: SpeechFake [14], a multilingual corpus covering 46 languages and 40 generation methods, and CodecFake [29], which focuses on neural codec-based forgeries. These datasets provide the diversity necessary to evaluate cross-lingual and cross-method generalization. Detailed statistics for both are provided in Appendix B.

Foundation Models for ADD. Pre-trained foundation models such as Whisper [30] have shown superior robustness in audio tasks due to extensive pre-training on 680k hours of diverse data. Recent research indicates that Whisper's self-attention architecture captures finer acoustic cues than CNN-based models do, improving cross-dataset generalization [31]. To further validate this, we conducted an extensive cross-subset evaluation on the CodecFake dataset (see Appendix C), which confirms that Whisper-based models effectively capture invariant spoofing patterns across diverse neural codec artifacts, establishing a robust foundation for the proposed AAFI and Mel-BAP modules.

Unlike existing works that typically treat Whisper as a fixed feature extractor [32], our framework integrates the AAFI module to enable dynamic, input-tailored compression. This addresses the

challenge of heterogeneous artifact densities across different forgeries, an aspect largely overlooked by static bottleneck architectures. Beyond representation learning, frequency-domain adversarial training has been explored to improve robustness in computer vision. FreqAT [33] combines adaptive frequency-domain transformations with adversarial training to improve model robustness while mitigating image quality degradation. Mask-constrained adversarial perturbations have also been explored in object detection [34]. In contrast, our Mel-BAP is designed specifically for deepfake detection. It incorporates the forensic insight that neural vocoder artifacts are disproportionately concentrated in high-frequency bands (4–8 kHz) [35], allowing for more targeted and efficient regularization than existing approaches.

3 Methodology

3.1 Overview

Let x denote an input audio waveform and y its corresponding ground-truth label. In this work, we adopt the Whisper architecture, which employs a fixed pre-processing module, denoted by $\mathcal{S}(\cdot)$, to convert raw waveforms into log-Mel spectrograms $M = \mathcal{S}(x)$. For ADD tasks, neural networks are typically trained to predict the label by minimizing the cross-entropy loss $\mathcal{L}(\hat{y}, y)$. Specifically, our proposed Mel-BAP strategy operates in the frequency domain by applying targeted perturbations directly to this log-Mel representation. Thus, the prediction is formulated as $\hat{y} = \mathcal{F}_\theta(\text{Mel-BAP}(M))$, and the model parameters θ are optimized by minimizing the following objective:

$$\mathcal{L}(\mathcal{F}_\theta(\text{Mel-BAP}(M)), y) \tag{1}$$

Justification for Decoder-based Classification. While a lightweight multilayer perceptron (MLP) head is standard for binary classification, we leverage the Whisper decoder for two strategic reasons. First, unlike MLP heads that typically require global pooling (sacrificing temporal granularity), the decoder's *cross-attention mechanism* dynamically attends to specific frames within the acoustic representation H_{acous}. This is critical for detecting deepfakes, where artifacts are often localized (e.g., phase discontinuities) rather than globally distributed. Second, utilizing the pre-trained decoder preserves the semantic-acoustic alignment learned during large-scale supervision, offering better initialization than using a randomly initialized classification head. The decoding task is simplified to generating a single class token (e.g., predicting the token for "Genuine" or "Spoofing" immediately after the start-of-transcript token), minimizing inference latency compared to full speech transcription.

Figure 1 illustrates the overall process, with an emphasis on the proposed Adaptive Audio Feature Integration (AAFI) module, which is positioned between the encoder and decoder components of the base model. During training, the entire model is optimized in an end-to-end manner. Crucially, to preserve the rich acoustic representations learned during large-scale pre-training and to minimize computational overhead, the parameters of the Whisper encoder remain fixed. The gradient updates are exclusively applied to the proposed AAFI module, the projection layers, and the Whisper decoder. To improve the granularity of the downstream classifier's input, we concatenate the shared features (the output of the base model encoder) with the selected features (the output of the proposed AAFI module) and pass them through a fully connected layer to adjust their dimensionality. The adaptive routing mechanism allows gradients to flow through the selected compression pools, enabling the adaptive router to learn meaningful selection policies based on audio characteristics. The Mel-BAP component further encourages the model to learn more robust and discriminative audio features. The detailed structure of Mel-BAP is presented in the left part of Figure 1. In the following sections, we provide a detailed description of the two core components of the proposed framework.

Unlike standard classification models that rely on a simple classification head, our framework leverages the generative capabilities of the Whisper decoder. First, we integrate the diverse information captured by the framework. Let $\mathbf{Z}_{enc} \in \mathbb{R}^{T \times d_{enc}}$ denote the shared features from the fixed Whisper encoder, and $\mathbf{Z}_{aafi} \in \mathbb{R}^{T \times d_{aafi}}$ denote the selected adaptive features from the AAFI module. To preserve both general acoustic patterns and detailed artifact information, we concatenate these features along the channel dimension: $\mathbf{Z}_{cat} = \text{Concat}(\mathbf{Z}_{enc}, \mathbf{Z}_{aafi})$. Subsequently, a fully connected (FC) layer is employed to map the concatenated features to the dimension required by the decoder (d_{model}): $\mathbf{H}_{acous} = \text{ReLU}(\mathbf{Z}_{cat}\mathbf{W}_{fc} + \mathbf{b}_{fc})$, where $\mathbf{W}_{fc}$ and $\mathbf{b}_{fc}$ are learnable parameters. $\mathbf{H}_{acous}$ serves as the robust acoustic representation fed into the Whisper decoder. The decoder operates in a sequence-to-sequence manner. It receives the acoustic representation $\mathbf{H}_{acous}$ (as keys and values

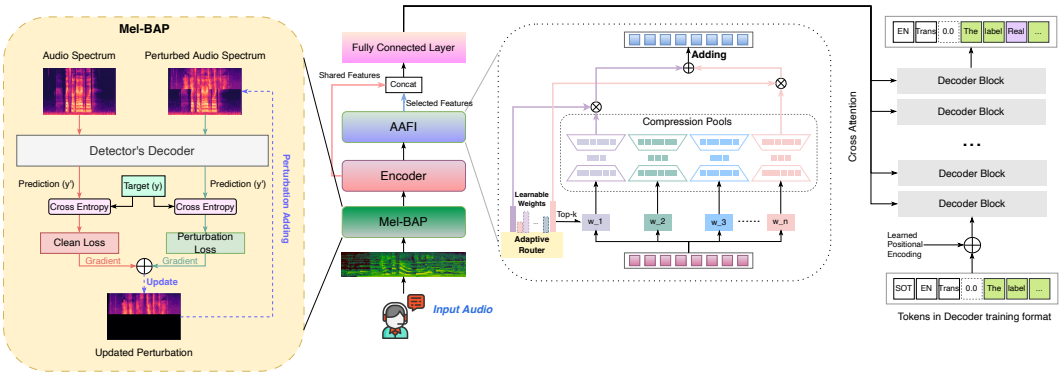

Figure 1: Schematic of the robust audio forgery detection pipeline. Left: Mel-BAP generates band-selective perturbations via PGD as targeted adversarial regularization. Center: AAFI implements an adaptive information bottleneck, using a router to dynamically select compressed features from heterogeneous pools. Right: Fusion of multi-resolution features for sequence-to-sequence decoding.

for cross-attention) and a sequence of prompt tokens (as queries). The model is trained to generate a target sequence $\mathbf{Y} = \{y_1, y_2, \ldots, y_L\}$ that represents the predicted label (e.g., the text tokens for "Genuine" or "Spoofing"). The probability of the target sequence is computed autoregressively:

$$P(\mathbf{Y}|\mathbf{H}_{acous}) = \prod_{t=1}^{L} P(y_t|y_{<t}, \mathbf{H}_{acous}) \tag{2}$$

In our ADD task, the prediction is determined by the probability assigned to the class-specific tokens in the generated sequence.

To optimize the framework end-to-end and handle data imbalance, we employ a weighted cross-entropy loss on the generated tokens. Let y_{target} be the ground-truth token index for a given sample, and $p(y_{target})$ be the probability predicted by the decoder for that token. The loss $\mathcal{L}_{WCE}$ is formulated as:

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p(y_{target}^{(i)}|\mathbf{H}_{acous}^{(i)}) \tag{3}$$

where N is the batch size, and w_{y_i} is the class weight. To ensure stable training and balanced gradient flow, we employed a hybrid approach combining oversampling with class weighting: we replicated the genuine samples (minority class) to reach an effective training ratio of 3:5 (Genuine:Spoofed). Accordingly, we set $w_{genuine} = 1.0$ and $w_{spoofed} = 0.6$ to prevent the model from becoming biased towards the majority class, thereby improving the detection of genuine speech (reducing the false rejection rate). Combined with the Mel-BAP strategy, the final objective minimizes the expected loss over the perturbed inputs:

$$\mathcal{L} = \mathcal{L}_{WCE}(\mathcal{F}_\theta(M), Y) + \mathcal{L}_{WCE}(\mathcal{F}_\theta(\text{Mel-BAP}(M)), Y) \tag{4}$$

Cross-lingual Generalization Analysis. Finally, to further stress-test the zero-shot transferability of our framework under extreme conditions, we evaluated its performance in five unseen languages (es, fr, ja, it, ko) from the SpeechFake MD dataset. Despite being trained only on bilingual (EN/CN) data, our model maintains robust detection capabilities across all linguistic boundaries. As detailed in Appendix A, these results provide additional evidence that the AAFI and Mel-BAP modules effectively capture language-invariant spoofing cues, preventing the model from relying excessively on language-specific acoustic patterns.

3.2 Adaptive Audio Feature Integration

Motivation. Audio deepfake detection (ADD) faces a fundamental challenge: the *heterogeneous information density* of spoofing artifacts. High-fidelity deepfakes (e.g., generated by advanced neural

codecs) exhibit subtle artifacts that require high-dimensional representations to capture, whereas low-quality spoofs often contain obvious noise, and high-dimensional features may introduce redundancy and overfitting in such cases. Existing methods typically employ fixed-size compression layers, creating a suboptimal trade-off between fidelity and compactness. To resolve this, we propose the **A**daptive **A**udio **F**eature **I**ntegration (**AAFI**) module. Unlike standard mixture-of-experts (MoE) architectures that utilize homogeneous experts to scale model capacity [36], AAFI acts as a *dynamic intrinsic filter* based on the information bottleneck principle. It adaptively selects the optimal compression ratio for each input sample, maximizing the mutual information between the latent representation and the spoofing labels while minimizing irrelevant noise.

Heterogeneous Compression Pool. To accommodate the varying information densities, we construct a heterogeneous compression pool consisting of C parallel branches. Unlike standard MoE experts that share identical architectures, each branch in our pool has a *distinct latent dimension* $d_i \in \{d_{\min}, \ldots, d_{\max}\}$. This design provides a spectrum of "resolutions" for the audio features. Given an encoded audio feature representation $\mathbf{Z}_{enc} \in \mathbb{R}^{T \times d_r}$ (where d_r equals the encoder's hidden dimension, e.g., 768 for Whisper-small), each branch performs dimensionality reduction via $\mathbf{Z}_i = \text{DownNet}_i(\mathbf{Z}_{enc}) \in \mathbb{R}^{T \times d_i}, \quad i = 1, \ldots, C$, where DownNet_i is implemented as a single-layer learnable linear projection to minimize computational overhead. Detailed architectural specifications, including the exact dimensionality settings and complexity analysis, are provided in Appendix D.

Adaptive Routing with Exploration. To dynamically assign the optimal information bottleneck for each sample, we first apply temporal mean pooling to the encoded features to obtain an utterance-level representation $\bar{z} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_{enc,t} \in \mathbb{R}^{d_r}$. We then employ an adaptive router to compute selection weights $w \in \mathbb{R}^C$. To facilitate exploration and prevent premature convergence to a suboptimal bottleneck, we inject input-dependent Gaussian noise into the routing mechanism. The selection mechanism is formulated as $w = \text{Softplus}(\bar{z}W_{router} + (\bar{z}W_{noise}) \odot \epsilon)$, where $W_{router} \in \mathbb{R}^{d_r \times C}$ and $W_{noise} \in \mathbb{R}^{d_r \times C}$ are learnable weight matrices, $\epsilon \sim \mathcal{N}(0, I)$ is a standard Gaussian noise vector, and $\odot$ denotes the element-wise product. This noise term acts as a stochastic regularizer that maintains routing diversity. By perturbing the routing logits, the model is forced to explore diverse compression resolutions during the initial training phase, ensuring that all branches in the pool are effectively optimized and preventing mode collapse. As demonstrated by the ablation study in Appendix E, this noise-driven exploration significantly enhances *branch utilization diversity* and improves generalization compared to deterministic routing.

Feature Integration. We employ a top-k selection strategy to balance efficiency and representational capacity. Let K be the set of indices corresponding to the k largest values in w. The adaptive feature representation is obtained by a weighted fusion of the upsampled features from the selected branches:

$$\mathbf{Z}_{aafi} = \sum_{i \in K} \frac{w_i}{\sum_{j \in K} w_j} \cdot \text{UpNet}_i(\mathbf{Z}_i) \tag{5}$$

where UpNet_i maps the compressed features $\mathbf{Z}_i$ back to the original dimension d_r. The adaptive feature representation $\mathbf{Z}_{aafi}$ is subsequently fused with the shared encoder features $\mathbf{Z}_{enc}$. The combined representation is then projected via a fully connected layer to form the final acoustic context H_{acous}, following the integration process described in Section 3.1.

3.3 Mel-Band Adversarial Perturbation

Motivation: Mitigating the Robustness-Accuracy Trade-off. State-of-the-art deepfake detectors often exhibit a pathological reliance on imperceptible high-frequency artifacts as discriminative cues [35]. While this yields exceptional performance on specific benchmarks, it renders models highly vulnerable to real-world distortions—such as lossy compression or environmental noise—that inevitably corrupt these fragile features. Although standard adversarial training or data augmentation can bolster robustness, they frequently incur a significant performance penalty on clean data, a dilemma recognized as the *robustness-accuracy trade-off* [37]. We posit that this degradation stems from the nature of global perturbations, which often act as an indiscriminate "blunt instrument" that obscures essential semantic content and compromises the global spectro-temporal integrity of the audio.

To bridge this gap, we propose the Mel-Band Adversarial Perturbation (Mel-BAP) strategy. Unlike conventional global attacks, Mel-BAP applies *band-selective perturbations* specifically targeting the mel-frequency regions where synthetic artifacts are most salient. This approach serves as a form of *targeted regularization*: by smoothing the decision boundary within sensitive feature subspaces, it prevents the model from exploiting "shortcuts" based on unstable high-frequency patterns. Consequently, Mel-BAP incentivizes the model to shift its focus toward *intrinsic, invariant representations* that characterize the underlying acoustic consistency. By constraining perturbations to localized bands, Mel-BAP effectively fortifies the model against both malicious adversarial attacks and heterogeneous real-world distortions without sacrificing the fidelity of global acoustic structures or clean-data accuracy.

Mechanism. Mel-BAP operates directly in the mel-spectrogram domain to align with the perceptual scaling of the Whisper encoder. Let $\mathbf{M} \in \mathbb{R}^{T \times F}$ denote the input mel-spectrogram derived from audio $\mathbf{x}$, where T is the number of time frames and $F = 80$ is the number of mel-frequency bins. We define a binary mask $\mathbf{D} \in \{0, 1\}^{T \times F}$ to restrict perturbations to a target mel-band $[m_l, m_u]$:

$$\mathbf{D}_{t,f} = \begin{cases} 1 & \text{if } f \in [m_l, m_u], \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

The adversarial perturbation $\boldsymbol{\delta}_m$ is optimized to maximize the detection loss $\mathcal{L}$ for the target model $\mathcal{F}_\theta$ within the constrained band. The perturbed input $\mathbf{M}'$ is computed as $\mathbf{M}' = \mathbf{M} + \mathbf{D} \odot \boldsymbol{\delta}_m$. We solve the maximization problem $\max_{\boldsymbol{\delta}_m} \mathcal{L}(\mathcal{F}_\theta(\mathbf{M} + \mathbf{D} \odot \boldsymbol{\delta}_m), y)$ subject to an L_∞-norm constraint $\|\boldsymbol{\delta}_m\|_\infty \leq \epsilon$ via projected gradient descent (PGD) [38]. Unlike standard adversarial training that often requires expensive parameter sweeps, we adopt a fixed configuration based on empirical stability:

$$\boldsymbol{\delta}_m^{(k+1)} = \Pi_\epsilon \Big(\boldsymbol{\delta}_m^{(k)} + \alpha \cdot \text{sign}\big(\nabla_{\boldsymbol{\delta}_m} \mathcal{L}(\mathcal{F}_\theta(\mathbf{M} + \mathbf{D} \odot \boldsymbol{\delta}_m^{(k)}), y) \big) \Big) \tag{7}$$

Specifically, we set the perturbation budget $\epsilon = 0.01$ (relative to the normalized spectral magnitude), the step size $\alpha = 0.002$, and the number of iterations $K = 5$. Π_ϵ projects the perturbation onto the ϵ-ball, and $\odot$ denotes element-wise multiplication. A detailed ablation study on the accuracy-cost trade-off associated with the number of iterations K is provided in Appendix F.

Mel-BAP shares the high-frequency regularization motivation of Frequency-Selective Adversarial Training (F-SAT, Zhang et al. [35]), while adapting the perturbation procedure to the Whisper pipeline: *1) Perceptual Alignment and Forensic Priors:* While F-SAT perturbs selected frequency bands in the STFT magnitude spectrum and reconstructs the waveform, Mel-BAP applies perturbations directly in Whisper's log-Mel input space. This aligns perturbation optimization with the representation consumed by the encoder. Mel-BAP targets high-frequency components to discourage reliance on fragile detection cues. Mel-BAP uses a fixed 4-8 kHz target band rather than random or sweeping band selection. We provide a frequency-band ablation in Table 3 to compare this constraint with full-band adversarial training. *2) End-to-End Computational Efficiency:* By modifying the log-Mel input directly, Mel-BAP avoids the waveform reconstruction and subsequent feature extraction required by the F-SAT pipeline, keeping perturbation optimization within the detector's input representation. *3) Zero Inference Cost:* Mel-BAP operates exclusively during training. It therefore adds no perturbation-generation overhead during inference.

4 Experimental Evaluation

Experimental Settings. Unless otherwise stated, our experiments primarily use the Bilingual Dataset (BD) from SpeechFake[1] [14], which contains high-fidelity deepfakes in English (EN) and Chinese (CN). The dataset includes samples generated using 30 open-source tools (e.g., TorToiSe, VITS, FastSpeech2) and 10 commercial APIs (e.g., Azure TTS, Google Wavenet), spanning text-to-speech (TTS), voice conversion (VC), and neural vocoder (NV) architectures. Based on the subset splits used in our evaluation, the BD dataset comprises over 1.19 million English samples and over 0.88 million Chinese samples, totaling more than 2 million high-fidelity deepfake and genuine utterances. We use the official SpeechFake BD train, dev, and test subsets, containing 75,708/12,618/37,854 real and 633,354/105,544/315,222 fake utterances, respectively. No samples

[1]Available from: `https://www.modelscope.cn/datasets/inclusionAI/SPEECHFAKE`

Table 1: Performance Comparison of Adaptive Routing (top-k) in the AAFI Module versus Whisper base. EER (%) results (lower is better).

Data Group	top-k number				Whisper-base
	top-2	top-3	top-4	top-5	
TTS	3.91	0.85	**0.83**	0.99	1.18
VC	4.85	1.88	**1.67**	2.11	3.32
NV	8.44	**3.44**	3.50	4.34	6.07
Average	5.73	2.06	**2.00**	2.48	3.52

are reassigned across the released subsets. An unseen test set (BD-UT) contains over 85,000 test samples, with all fake samples generated exclusively using commercial APIs unseen during training. Detailed dataset statistics are provided in Table 8 in Appendix B.

All experiments were conducted on two NVIDIA A6000 GPUs. We optimized the model using AdamW [37] with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. To address class imbalance, we employed a weighted cross-entropy loss with weights of 1.0 for genuine samples and 0.6 for spoofed samples. Training used a batch size of 32 for up to 10 epochs, with early stopping triggered if the validation EER did not improve for five consecutive epochs. Inputs were 16 kHz raw waveforms. We evaluated performance using the widely adopted EER metric. Our base models were Whisper-base (74M parameters) and Whisper-small (244M parameters).

To apply Mel-BAP, the audio input $\mathbf{x}$ is first converted into a log-Mel spectrogram $\mathbf{M} \in \mathbb{R}^{T \times F}$ (where $F = 80$ is the number of mel bins) using Whisper's default preprocessing pipeline (25 ms window, 10 ms hop, 80 mel bins, log-compression). A target Mel-bin range $[m_l, m_u]$ is then selected based on artifact analysis (e.g., $m_l = 60$ and $m_u = 80$). Within this band, adversarial perturbations $\boldsymbol{\delta}_m$ are generated using five PGD iterations and added to obtain the perturbed spectrogram $\mathbf{M}' = \mathbf{M} + \mathbf{D} \odot \boldsymbol{\delta}_m$, which is subsequently fed into Whisper's encoder for prediction.

4.1 AAFI Ablation and Routing Analysis

We conduct targeted experiments on the SpeechFake dataset to validate the core functionality and design choices of the AAFI module. These experiments focus on two key aspects: (1) whether AAFI dynamically adapts to input characteristics as intended, and (2) how to select the optimal value of k for top-k routing. All AAFI-only variants are trained on the BD training set under identical conditions. We evaluate them on the three SpeechFake BD validation subsets (TTS, VC, and NV). For each sample, we record the compression dimensions (16/32/64/128/256) selected by AAFI's adaptive router. To reduce computational cost and ensure a fair comparison, we freeze all encoder layers across all evaluated configurations (including the Whisper-base baseline).

Table 1 compares the performance of the adaptive routing strategy within the AAFI module under different top-k configurations against the base model without AAFI. Incorporating AAFI with a sufficient routing capacity ($k \geq 3$) results in a clear improvement in EER across all validation groups. Conversely, excessively restricting the feature integration to top-2 degrades performance below the baseline (5.73% vs. 3.52% in average EER), indicating that sufficient diversity in compression resolutions is essential. Although top-4 achieves the best performance (2.00% in average EER), the difference between top-3 and top-4 is minimal (0.85% vs. 0.83% for TTS, 1.88% vs. 1.67% for VC, 3.44% vs. 3.50% for NV, and 2.06% vs. 2.00% on average), suggesting that top-3 already provides nearly optimal performance. Extending to top-5 slightly reduces performance, likely due to the inclusion of redundant information, while increasing computational cost. Overall, top-3 offers the best trade-off between efficiency and accuracy and is the most practical choice for deployment. Furthermore, a stratified analysis shows that routing distributions vary across spoofing conditions: NV samples assign 59% of their selections to the wider 128/256-d branches, whereas TTS samples assign 64% to the narrower 16/32-d branches. The observed distributions are consistent with condition-dependent multi-resolution selection; full results are provided in Appendix H.

Table 2: Ablation study on the effectiveness of Mel-BAP across SpeechFake BD subsets. Lower EER (%) indicates better performance.

Model	BD-EN	BD-CN	BD-UT	Average
w/o Mel-BAP	1.94	1.68	1.19	1.60
Mel-BAP Only	1.35	1.20	0.91	1.14
AAFI + Mel-BAP	**1.01**	**0.82**	**0.53**	**0.79**

4.2 Mel-BAP Ablation Study

Module Ablation. We evaluate Mel-BAP's efficacy by comparing *1) Whisper base*, *2) Base+Mel-BAP*, and *3) AAFI+Mel-BAP* across SpeechFake subsets. Table 2 shows that Mel-BAP consistently reduces EER (e.g., from 1.19% to 0.91% on BD-UT), demonstrating its ability to stabilize predictions against unseen spoofing patterns. Integrating AAFI further lowers EER to 0.53%, confirming that adaptive feature integration complements targeted perturbations by selecting optimal representations for different inputs.

Frequency-Band Ablation. To examine the effect of the frequency-band constraint, we compare high-, full-, and low-band perturbations using Whisper-base without AAFI or standard augmentation. All variants use the same perturbation budget ($\epsilon = 0.01$, $\alpha = 0.002$, $K = 5$). Table 3 shows that Mel-BAP achieves an EER of 0.91%, compared with 1.15% for full-band PGD and 1.28% for low-band PGD.

4.3 Robustness under Real-World Distortions

To assess real-world resilience, we subjected the BD-UT set to dynamic attacks, including background noise (NoiseX-92 [39]), speed adjustments (0.5×–1.5×), and volume fluctuations (±15 dBFS). The results are reported in Table 4. *First,* Mel-BAP significantly enhances robustness across all tested distortions. For both Whisper-base and Whisper-small, it consistently reduces EER in both clean and perturbed conditions (e.g., lowering EER under dynamic attacks from 3.57% to 2.44% for Whisper-base). *Second, and most critically,* Mel-BAP demonstrates a decisive advantage over standard global audio augmentation. As shown in the "Global Aug" columns, *even when the model is trained with the identical types of perturbations used in our evaluation (i.e., the same noise, speed, and volume variations)*, simply applying them as indiscriminate global distortions during training significantly degrades performance. This leads to substantial performance degradation even on clean data (e.g., EER rising from 1.19% to 2.33% for Whisper-base). This result highlights a classic robustness-accuracy trade-off, in which global perturbations may inadvertently destroy essential semantic and structural cues. In sharp contrast, Mel-BAP's band-selective strategy functions as a *targeted regularizer*. By forcing the model to ignore fragile high-frequency artifacts while preserving the overall acoustic structure, it encourages the extraction of intrinsic, invariant features rather than overfitting to noise patterns. *Finally,* while larger models (Whisper-small) provide stronger baselines, the relative improvement from Mel-BAP is more pronounced in Whisper-base. This suggests that specialized mechanisms like Mel-BAP are crucial for preventing shortcut learning and sustaining generalization in non-stationary, real-world scenarios.

Beyond standard real-world distortions, we evaluate the adversarial robustness of our framework against strong test-time adaptive attacks (including standard PGD, CW, and frequency-aware Adaptive PGD). Detailed evaluation protocols and results are provided in Appendix G.

4.4 Comparison with State-of-The-Art Methods

Fairness of Comparison and Zero-shot Protocol. A strictly fair comparison with state-of-the-art (SOTA) systems is challenging due to the diversity of proprietary training data used in commercial models (e.g., DLMSL, Resemble) and varying baselines. Therefore, our evaluation focuses on *zero-shot generalization potential* in real-world scenarios rather than algorithmic superiority under controlled training conditions. Results for the 25 external systems in Table 5 are taken from Speech-DF-Arena 2025 [40]. Because these systems use heterogeneous or undisclosed training protocols, the

Table 3: Ablation of adversarial frequency band constraints on the BD-UT dataset (Clean EER).

Strategy	Frequency Band	Mel-bin Indices	Clean EER (%)
(a) **Mel-BAP (Ours)**	4–8 kHz	[60, 80]	**0.91**
(b) Full-band PGD	0–8 kHz	[0, 80]	1.15
(c) Low-band PGD	0–4 kHz	[0, 60]	1.28

Table 4: Robustness comparison against standard augmentation under real-world distortions. EER (%) is reported on BD-UT.

Model	w/o Mel-BAP		with Mel-BAP		Global Aug	
	Clean	Attack	Clean	Attack	Clean	Attack
Base	1.19	3.57	**0.91**	**2.44**	2.33	4.68
Small	0.95	3.18	**0.87**	**2.25**	1.65	4.01

table provides a contextual rather than strictly controlled comparison. Our four Whisper-based models were trained on the same official SpeechFake BD splits and evaluated under the same zero-shot protocol. The term "zero-shot" here refers to evaluation on datasets (e.g., ADD 2022, In-the-Wild) not used during training.

To ensure a comprehensive evaluation across diverse and realistic scenarios, we evaluated our models on FakeOrReal [52] and its second version FakeOrReal-norm, In-the-Wild [13], the Audio Deepfake Detection Challenge (ADD 2022) track 1 evaluation set [11], ASVspoofing 2019 [9] and ASVspoofing 2021 LA [12]. FakeOrReal and In-the-Wild are English-language datasets that are widely used in the deepfake detection community to assess a detector's generalization performance. ADD 2022 is a Mandarin-language challenge dataset, and its deepfake speech samples were collected from the generation task of the same challenge. This selection covers the full spectrum of modern deepfake threats, ranging from high-fidelity lab-generated samples to unconstrained real-world recordings, thereby surpassing the limitations of traditional controlled benchmarks. In our study, we used only the audio portion, which includes samples in more than 50 languages. We used only the test partitions of the selected test datasets; their training portions, if available, were excluded entirely from both the model training and testing.

Table 5 presents a *comprehensive zero-shot evaluation* against 25 state-of-the-art systems and classic baselines. The results reveal several critical insights into the current landscape of deepfake detection. *First,* while certain specialized models exhibit exceptional performance on specific benchmarks, they often suffer from a "generalization collapse" in adverse environments. For instance, MoLEx achieves near-perfect results on *FakeOrReal* (0.18%) and *In-the-Wild* (0.03%), yet its performance catastrophically degrades to 31.94% on ADD 2022 Track 1. Similarly, the commercial leader DLMSL-SpeakSure-v0.1 dominates the clean *ASV2019 and ASV2021 LA* benchmarks ($< 0.1\%$) but struggles with the complex distortions in ADD 2022 (15.14%). This discrepancy suggests that these models may rely on "shortcuts", namely fragile high-frequency artifacts that are easily obliterated by real-world noise or lossy compression. *Second,* our proposed *Whisper-small + AAFI + Mel-BAP* model (252M parameters) emerges as the most robust generalist, delivering the lowest overall *average EER of* 1.51%. Most notably, it achieves a SOTA EER of 1.87% on ADD 2022 Track 1, outperforming the 3-billion-parameter Resemble-Detect-3B-Omni (9.96%) by a factor of five despite using approximately $11.9\times$ fewer parameters. This demonstrates that Mel-BAP effectively incentivizes the model to capture intrinsic acoustic consistency rather than overfitting to dataset-specific noise patterns. *Third,* the proposed approach maintains a superior efficiency-performance Pareto frontier. With only 252M parameters, it achieves a 76.3% lower average EER than the DF_Arena_1B_V_1 with 1 billion parameters (1.51% vs. 6.37%) and is significantly more accurate than massive commercial models. The integration of the AAFI + Mel-BAP modules allows the detector to generalize to novel forgery techniques and dynamic attack scenarios without fine-tuning.

Table 5: Comprehensive Zero-shot EER (%) results across 29 systems. All models are evaluated without fine-tuning. The Average is recalculated based on five key benchmarks: In-the-Wild, ASV2019, ASV2021 LA, FakeOrReal, and ADD 2022 Track 1. Values in bold indicate the best results per column. Lower EER indicates superior performance.

Model ID	**# Params**	**ADD 2022** (T1)	**FakeOrReal**	**In-the-Wild**	**ASV2019**	**ASV2021 LA**	**Average**
Literature and Commercial SOTA (Speech-DF-Arena 2025)							
Resemble-Detect-3B-Omni [40]	3,000 M	9.96	2.34	1.35	2.53	3.02	3.84
DLMSL-SpeakSure-v0.1 [40]	658.6 M	15.14	0.27	1.28	**0.04**	**0.08**	3.36
Whispeak [40]	98.9 M	11.94	1.02	1.27	0.39	3.58	3.64
DF_Arena_1B_V_1 [41]	1,000 M	22.21	2.92	0.91	1.14	4.66	6.37
Momenta [40]	350 M	25.47	2.87	4.63	1.21	4.88	7.81
Syntra Detector [40]	584 M	23.58	0.53	3.98	1.48	14.05	8.73
DF_Arena_500M_V_1 [40]	500 M	23.98	2.30	1.76	1.09	4.23	6.67
MoLEx [40]	376.4 M	31.94	**0.18**	**0.03**	0.28	6.33	7.75
Resemble Detect [40]	2,112 M	28.22	1.37	3.95	1.32	1.65	7.30
DF_Arena_100M_V_1 [40]	100 M	27.06	7.42	2.22	1.54	7.61	9.17
DF_Arena_100M_V_0 [40]	100 M	31.96	6.14	4.29	4.16	10.06	11.32
XLSR + SLS [16]	340 M	33.95	5.08	7.46	0.23	2.87	9.92
TCM [42]	319 M	37.40	10.69	7.79	0.19	3.00	11.81
BiCrossMamba-ST [43]	318.2 M	30.44	6.85	7.94	0.71	3.83	9.95
Nes2NetX [44]	317.9 M	34.47	6.32	7.75	0.12	2.17	10.17
Wav2Vec2 AASIST [45]	317.8 M	31.05	7.47	11.20	0.22	0.82	10.15
XLSR Mamba [46]	319 M	34.23	6.71	6.71	0.42	0.93	9.80
Whisper Mesonet [32]	7.6 M	38.38	47.75	26.73	5.83	15.82	26.90
Wav2Vec2 ECAPA [47]	324 M	46.44	62.32	30.70	29.70	26.61	39.15
AASIST [48]	0.3 M	51.06	21.64	43.01	0.83	11.46	25.60
WavLM ECAPA [47]	102 M	44.17	23.37	34.65	0.76	6.68	21.92
RawGatST [49]	0.4 M	50.00	53.09	52.54	1.06	10.25	33.39
RawTFNet [50]	0.2 M	43.85	36.57	38.73	1.89	5.04	25.22
Hubert ECAPA [47]	102 M	47.66	46.22	38.66	1.06	12.55	29.23
Rawnet2 [51]	17.6 M	49.36	65.68	49.19	33.04	40.07	47.47
Our Proposed Models							
Whisper-base w/o (AAFI+Mel-BAP)	74 M	3.11	0.91	5.53	1.94	2.15	2.73
Whisper-base + (AAFI+Mel-BAP)	80 M	2.16	0.55	4.78	1.19	1.30	2.00
Whisper-small w/o (AAFI+Mel-BAP)	244 M	2.94	0.73	4.36	1.10	1.77	2.18
Whisper-small + (AAFI+Mel-BAP)	252 M	**1.87**	0.49	3.14	0.93	1.12	**1.51**

5 Conclusion

This paper presents a robust and scalable framework for generalized ADD by integrating AAFI and Mel-BAP into a Whisper-based architecture. Our approach effectively bridges the gap between representation fidelity and adversarial robustness. By leveraging a stochastic compression pool and targeted mel-domain regularization, the framework achieves superior zero-shot generalization across diverse languages and synthesis domains. Experimental results confirm that our model establishes a new efficiency-performance Pareto frontier for ADD, offering a modular "plug-in" solution that can be seamlessly integrated into existing security pipelines. Future work will focus on further refining adaptive mechanisms to support scalable real-time detection in non-stationary acoustic environments.

Acknowledgements

This research was supported in part by the Self-initiated Project of Yangtze Delta Region Institute of Tsinghua University, Zhejiang No. 140104370. All authors are grateful to the anonymous referees for their valuable comments that led to considerable improvements to this article.

References

[1] M. Li, Y. Ahmadiadli, and X. P. Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38, 2025.

[2] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, and J. Tao. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 111(10):1355–1381, 2023.

[3] Z. Ye, Z. Ju, H. Liu, X. Tan, J. Chen, Y. Lu, and Y. Guo. Flashspeech: Efficient zero-shot speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6998–7007, October 2024.

[4] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

[5] Yi Ren, Yang Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[6] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1024–1037, 2020.

[7] T. Zhang, L. Deng, L. Zhang, and X. Dang. Deep learning in face synthesis: A survey on deepfakes. In *Proceedings of the IEEE 3rd International Conference on Computer Communication and Engineering Technology (CCET)*, pages 67–70, 2020.

[8] H. Wu, J. Kang, L. Meng, H. M. Meng, and H.-Y. Lee. The defender's perspective on automatic speaker verification: An overview. In *Proceedings of the Workshop on Deepfake Audio Detection and Analysis (DADA) at IJCAI*, 2023.

[9] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, et al. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of Interspeech*, pages 1008–1012, 2019.

[10] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, et al. 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.

[11] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, et al. Add 2022: The first audio deep synthesis detection challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220, 2022.

[12] 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. In *Proceedings of the ASVspoof 2021 Workshop on Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[13] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Proceedings of Interspeech*, pages 2783–2787, 2022.

[14] Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Speechfake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9985–9998, 2025.

[15] Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[16] Q. Zhang, S. Wen, and T. Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6765–6773, October 2024.

[17] Jianqiao Cui, Bingyao Yu, Qihao Wang, Fei Meng, and Jiwen Lu. Whiadd: Semantic-acoustic fusion for robust audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11610–11618, 2025.

[18] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019.

[19] Kenji Kawaguchi, Zhun Deng, Xiang Ji, and Jian Huang. How does information bottleneck help deep learning? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 16049–16096. PMLR, July 2023.

[20] Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *NeurIPS*, 2023.

[21] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference (WWW)*, pages 4095–4106, 2024.

[22] Juan M. Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9241–9245. IEEE, 2022.

[23] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12702–12706. IEEE, 2024.

[24] Yanghao Zhang, Wenjie Ruan, Fu Wang, and Xiaowei Huang. Generalizing universal adversarial perturbations for deep neural networks. *Machine Learning*, 112(5):1597–1626, 2023.

[25] Fu Wang, Yanghao Zhang, Yanbin Zheng, and Wenjie Ruan. Dynamic efficient adversarial training guided by gradient magnitude. *arXiv preprint arXiv:2103.03076*, 2023.

[26] Fu Wang, Zeyu Fu, Yanghao Zhang, and Wenjie Ruan. Self-adaptive adversarial training for robust medical segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 725–735. Springer, 2023.

[27] Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24746–24755, June 2024.

[28] Feng Dong, Liu Wang, Xu Nie, Fei Shao, Haoyu Wang, Ding Li, Xiapu Luo, and Xusheng Xiao. Distdet: A cost-effective distributed cyber threat detection system. In *Proceedings of the USENIX Security Symposium*, 2023.

[29] Haibin Wu, Yuan Tseng, and Hung yi Lee. Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2024.

[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[31] Qian Luo and Kalyani Vinayagam Sivasundari. Whisper+ aasist for deepfake audio detection. In *International Conference on Human-Computer Interaction*, pages 121–133. Springer, 2024.

[32] Piotr Kawa, Marcin Plata, Michal Czuba, Piotr Szymanski, and Piotr Syga. Improved deepfake detection using whisper features. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4009–4013, Dublin, Ireland, 2023. ISCA.

[33] Denghui Zhang, Yanming Liang, Qiangbo Huang, Xvxin Huang, Peixin Liao, Ming Yang, and Liyi Zeng. Freqat: an adversarial training based on adaptive frequency-domain transform. In *International Conference on Advanced Data Mining and Applications*, pages 287–301. Springer, 2024.

[34] Yanghao Zhang, Fu Wang, and Wenjie Ruan. Fooling object detectors: Adversarial attacks by half-neighbor masks. *arXiv preprint arXiv:2101.00989*, 2021.

[35] Zirui Zhang, Wei Hao, Aroon Sankoh, William Lin, Emanuel Mendiola-Ortiz, Junfeng Yang, and Chengzhi Mao. I can hear you: Selective robust training for deepfake audio detection. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Representation Learning*, volume 2025, pages 9728–9746, 2025.

[36] V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, and S. Tubaro. Leveraging mixture of experts for improved speech deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, April 2025.

[37] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[39] Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251, 1993.

[40] Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Matthew Magimai Doss. Speech DF Arena: A Leaderboard for Speech DeepFake Detection Models. `https://arxiv.org/abs/2509.02859`, 2025.

[41] Ajinkya Kulkarni, Atharva Kulkarni, Sandipana Dowerah, Matthew Magimai Doss, and Tanel Alumäe. DF_Arena_1B_V_1 - Universal Audio Deepfake Detection, 2025. Accessed: 2025-12-24.

[42] 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 *Proc. INTERSPEECH 2024*, pages 402–406, 2024.

[43] Yassine El Kheir, Tim Polzehl, and Sebastian Möller. BiCrossMamba-ST: Speech Deepfake Detection with Bidirectional Mamba Spectro-Temporal Cross-Attention. In *Proc. INTERSPEECH 2025*, 2025. Forthcoming.

[44] Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. Nes2Net: A Lightweight Nested Architecture for Foundation Model Driven Speech Anti-Spoofing. *IEEE Transactions on Information Forensics and Security*, 20:12005–12018, 2025.

[45] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas W. D. Evans. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 119–126, 2022.

[46] Yang Xiao and Rohan Kumar Das. XLSR-Mamba: A Dual-Column Bidirectional State Space Model for Spoofing Attack Detection. *IEEE Signal Processing Letters*, 32:1276–1280, 2025.

[47] Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, and Mathew Magimai Doss. Exploring generalization to unseen audio data for spoofing: Insights from ssl models. In *ASVSpoof workshop 2024*, 2024.

[48] J. w. Jung, H.-S. Heo, H. Tak, H. j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. ICASSP 2022*, pages 6367–6371, 2022.

[49] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 1–8, 2021.

[50] Yang Xiao, Ting Dang, and Rohan Kumar Das. Rawtfnet: A lightweight cnn architecture for speech anti-spoofing. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1997–2001. IEEE, 2025.

[51] 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.

[52] R. Reimao and V. Tzerpos. FoR: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10, 2019.

[53] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France, 2020. European Language Resources Association (ELRA).

[54] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pages 271–350, 2019.

Table 6: Cross-lingual evaluation results on unseen languages (SpeechFake MD). EER (%) is compared between the baseline and our proposed method.

Language	Whisper-base EER (%)		Improv. ↓
	w/o AAFI+BAP	w/ AAFI+BAP	
Spanish (es)	7.11	**5.65**	1.46
French (fr)	5.17	**3.92**	1.25
Japanese (ja)	5.52	**4.19**	1.33
Italian (it)	4.66	**3.91**	0.75
Korean (ko)	4.50	**4.02**	0.48
Average	**5.39**	**4.34**	**1.05**

A Ablation Study on Unseen Languages

To assess the cross-lingual generalization ability of our proposed model in the ADD task, we conduct an ablation study on the SpeechFake Multilingual dataset (MD) [14]. The MD covers a total of 46 languages, including 9 primary languages—English (en), Chinese (zh), Spanish (es), French (fr), Hindi (hi), Japanese (ja), Korean (ko), Persian (fa), and Italian (it)—together with 37 additional languages. Real speech samples in MD are sourced from the CommonVoice corpus [53], ensuring broad multilingual coverage.

For this experiment, we evaluate the **Whisper-base** model with both AAFI and Mel-BAP. The training data are drawn exclusively from the SpeechFake BD training subset, which contains only Chinese and English speech, thereby simulating a limited bilingual training scenario. During testing, we focus on a representative set of unseen languages, including Spanish (es), French (fr), Japanese (ja), Italian (it), and Korean (ko). For each language, we randomly sample 10,000 utterances from the MD for evaluation. Notably, the samples from the MD share nearly no overlap with BD. Generalization to MD is therefore more challenging than to BD validation subsets. This setup allows us to rigorously examine the model's robustness to cross-lingual transfer, particularly under conditions where the training and testing languages are entirely disjoint.

Table 6 reports the cross-lingual evaluation results on five unseen languages from the SpeechFake MD dataset. Several key observations can be made:

- First, across all languages, the Whisper-base model equipped with AAFI and Mel-BAP consistently outperforms the baseline without these components, achieving lower EER values. This confirms that the proposed strategies not only improve detection accuracy in in-domain conditions but also enhance robustness under cross-lingual settings where the model encounters unseen languages.
- Second, the degree of improvement varies by language. The most notable gains are observed for Spanish and French, where the integration of AAFI and Mel-BAP reduces the EER by approximately 1.5% and 1.2%, respectively. This suggests that these modules effectively capture language-invariant spoofing cues, enabling better generalization across linguistic boundaries. For other languages such as Japanese, Italian, and Korean, the improvements remain consistent but slightly smaller, indicating that language-specific acoustic and phonetic properties may affect the transferability of spoofing detection features.
- Finally, the results reinforce the effectiveness of Whisper as a baseline model for ADD. Thanks to Whisper's original powerful multilingual transcription ability, when combined with AAFI and Mel-BAP, the proposed model not only leverages its strong transcription and representation learning capabilities but also demonstrates resilience against cross-lingual distribution shifts.

B Dataset Statistics

CodecFake [29] (shown in Table 7), derived from the VCTK corpus [54], focuses on neural codec-based synthesis. It contains over one million samples generated via various codec frameworks, providing a rigorous benchmark for detecting increasingly prevalent codec-based forgeries.

Table 7: Details of adopted neural audio codec models. $A - F$ represents different codec frameworks. N_c represents the number of codebooks, SR represents audio sampling rate (in kHz), and BPS represents bits per second.

ID	Sub-ID	Model	Name	Training Data	N_c	SR	BPS
A	A1	SpeechTokenizer	16k	Librispeech	8	16	4
B	B1	HiFi-Codec	hifi_16k_320d	LibriTTS	4	16	2
	B2		hifi_16k_320d_large_uni	VCTK	4	16	2
	B3		hifi_24k_320d	AISHELL	4	24	3
C	C1	AudioDec	24k_320d	Valentini	8	24	6.4
D	D1	EnCodec	16k	Common Voice, DAPS	12	16	6
	D2		24k	MUSDB, VCTK	32	24	24
	D3		44k	AudioSet, Jamendo	9	44.1	8
E	E1	DAC	24k	Common Voice, DAPS, Jamendo, AudioSet, FSD50K	8	24	1.5
F	F1	FunCodec	en_libritts_16k_gr1nq32ds320	LibriTTS Subset	32	16	16
	F2		en_libritts_16k_gr8nq32ds320		32	16	16
	F3		en_libritts_16k_nq32ds320		32	16	16
	F4		en_libritts_16k_nq32ds640		32	16	8
	F5		zh_en_16k_nq32ds320	In-house, 25k hrs (en & zh-cn)	32	16	16
	F6		zh_en_16k_nq32ds640		32	16	8

Table 8 presents the detailed statistics of the SpeechFake Bilingual Dataset (BD) subsets used in our experiments.

Table 8: Statistics of SpeechFake Bilingual Dataset (BD) Subsets.

Subset	Content	Number of Samples	
		Real	Fake
BD-EN	English-only	64,000	1,128,882
BD-CN	Chinese-only	62,180	824,134
BD-UT	Unseen Techniques	37,854	50,000
BD-TTS	TTS-generated	126,180	1,014,107
BD-VC	VC-generated	126,180	535,205
BD-NV	NV-generated	126,180	403,704

C Robustness and Generalization in Cross-Subset Testing Based on Whisper

Cross-testing is essential to evaluate whether the model can generalize across diverse audio inputs derived from different types. This setting provides a robust assessment of the model's adaptability and consistency under varying data distributions. In this ablation study, we used the CodecFake dataset for the experiment, where the dataset was split into training (80%), validation (10%), and test (10%) sets, independently. All cross-testing experiments were conducted using the Whisper-base model (labelled "ADD whisper") on the CodecFake dataset, with results illustrated in Figure 2. The figure presents the Equal Error Rate (EER) across different training and testing subsets of the CodecFake dataset. Each row corresponds to a distinct training subset, while each column represents a testing subset, forming a matrix-like visualization that comprehensively reflects performance across subset combinations. The key findings are as follows:

The model trained on the full CodecFake dataset (row 1) achieves consistently lower EER values across nearly all test subsets, demonstrating strong generalization across diverse spoofing conditions. This advantage stems from the model's exposure to a wide distribution of codec-induced variations, which enables it to capture richer spoofing patterns and reduces the risk of overfitting to narrow conditions. In contrast, models trained on individual subsets (rows A–F6) tend to specialize in the

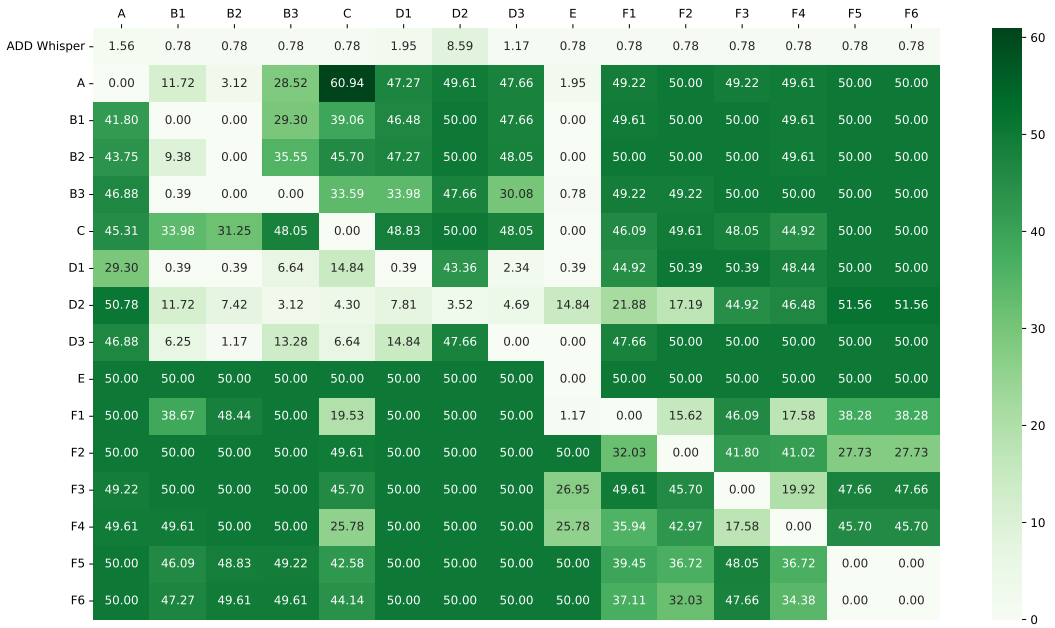

Figure 2: Cross-testing results of anti-spoofing models on CodecFake test subsets. We report equal error rate (EER), where lower is better. Rows A-F6 are ADD whisper models trained on one training subset of CodecFake. Columns A-F6 indicate the specific CodecFake test subset evaluated.

corresponding codec characteristics and thus achieve relatively low EERs on their in-distribution counterparts (e.g., row A $\rightarrow$ column A or row B1 $\rightarrow$ columns B1–B3). However, their performance degrades markedly under out-of-distribution testing, where EER values increase due to the lack of exposure to unseen codec artifacts. This contrast highlights a key finding: equipping the proposed model with more comprehensive knowledge of spoofing patterns yields consistently lower error rates and superior robustness.

Furthermore, the effectiveness of Whisper in this setting underscores its validity as a strong baseline for audio deepfake detection. Owing to its large-scale pre-training on diverse multilingual and multi-domain speech data, Whisper provides robust acoustic and semantic representations that transfer effectively to anti-spoofing tasks. When fine-tuning with structured prompts, as in ADD Whisper, the model not only preserves its transcription capability but also learns to associate subtle semantic–acoustic inconsistencies with spoofing cues. This dual capability makes Whisper particularly well-suited for evaluating and benchmarking ADD methods. Taken together, our results confirm that combining a powerful pre-trained backbone with comprehensive training data is critical for building reliable and generalizable deepfake detection systems.

D Implementation Details of AAFI

D.1 Architectural Specifications and Complexity

To ensure that the performance gains stem from adaptive feature selection rather than increased model capacity, the AAFI module is designed to be lightweight.

- **Dimensionality (d_r):** The input dimension d_r corresponds strictly to the hidden size of the backbone encoder: $d_r = 512$ for Whisper-base and $d_r = 768$ for Whisper-small.
- **Projection Layers:** Both DownNet_i and UpNet_i consist of a single linear transformation (fully connected layer) without bias, which intrinsically imposes the information bottleneck by projecting the features to the reduced latent dimension.
- **Parameter Overhead:** The total number of parameters introduced purely by the heterogeneous compression pool (with latent dimensions $d_i \in \{16, 32, 64, 128, 256\}$) is approximately $2 \times d_r \times \sum d_i$, which amounts to < 0.8 million parameters for Whisper-small.

However, the comprehensive parameter overhead reported in Table 5 (+6M for Whisper-base and +8M for Whisper-small) accounts for all additional trainable components integrated with the frozen backbone. This primarily originates from the fully connected fusion layer (W_{fc}) that projects the concatenated features ($\mathbf{Z}_{cat} \in \mathbb{R}^{T \times 2d_r}$) back to the decoder dimension d_{model}, alongside the parameters in the adaptive routing module. Despite this, the total added capacity remains marginal (e.g., approximately 3.2% increase for Whisper-small), substantiating our claim that the observed generalization gains are driven by the adaptive mechanism rather than simple parameter scaling.

D.2 Router Noise Schedule and Stability Analysis

The stochastic exploration mechanism is critical for preventing the router from collapsing into a trivial solution (i.e., always selecting the largest bottleneck).

- **Noise Schedule:** We employ a constant noise injection scheme during training where the standard Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is scaled by an input-dependent learned variance, specifically $(\bar{z}W_{noise}) \odot \epsilon$. During inference, we set $\epsilon = 0$ to ensure deterministic and reproducible predictions.
- **Top-k Selection Stability:** As shown in Table 1 of the main paper, we empirically found that $k = 3$ offers the optimal trade-off. To further validate stability, we analyzed the entropy of the routing weights w throughout training. The injection of noise maintains high entropy in the early epochs, encouraging exploration of all compression branches, before gradually settling into a stable selection policy driven by the audio characteristics.

E The Role of Stochasticity in Feature Integration.

Table 9: Ablation study of Gaussian noise in AAFI on the BD-UT dataset.

Configuration	EER (%)	Utilization Entropy ↑
Baseline (w/o AAFI)	1.19	-
AAFI w/o Noise	1.05	0.74
AAFI w/ Noise	**0.91**	**0.92**

To address the necessity of the Gaussian noise introduced in the AAFI module, we conduct an ablation study on the BD-UT dataset, comparing the model's performance with and without this stochastic component. As shown in Table 9, the inclusion of Gaussian noise ($\epsilon \sim \mathcal{N}(0, I)$) yields a consistent improvement in EER (e.g., from 1.05% to 0.91% on the test set), compared to the 1.19% EER of the frozen baseline.

Theoretically, without noise, the routing mechanism is susceptible to *representation collapse*, where the model prematurely converges to a deterministic sub-optimal combination of features—typically over-relying on the most dominant pre-trained layers. By introducing a stochastic perturbation during training, we facilitate **active exploration** of the feature manifold.

This ensures that the router does not settle into a "winner-take-all" state but instead maintains routing diversity. Specifically, we calculate the Expert Utilization Entropy H to quantify this diversity. Let P_i denote the normalized, empirical utilization frequency of the i-th branch across the evaluation set, derived from the routing weights. The entropy is mathematically defined as:

$$H = -\sum_{i=1}^{C} P_i \log_2(P_i) \tag{8}$$

Our results show that the inclusion of noise leads to a significantly higher entropy ($H = 0.92$ vs. $H = 0.74$), indicating that the model is forced to explore and utilize a broader range of spectro-temporal features. This diverse integration is precisely what enables the model to maintain high accuracy across heterogeneous datasets.

F Effect of PGD Iteration Number in Mel-BAP

To better understand the accuracy-cost trade-off introduced by the Mel-BAP strategy, we further analyzed the effect of the PGD iteration number (K) on the BD-UT dataset using the Whisper-small encoder. The resulting EERs and estimated training costs (relative to the baseline $K = 0$) are reported in Table 10.

Table 10: Ablation study on the effect of PGD iteration number (K) in Mel-BAP on the BD-UT dataset.

Number of PGD iterations	**EER (%)**	**Estimated Training Time Increase**
$K = 0$ (Original in Table 4)	0.95	-
$K = 1$	1.02	$\sim$25%
$K = 3$	0.89	$\sim$70%
$K = 5$ (Our Setting)	**0.87**	$\sim$115%
$K = 10$	**0.86**	$\sim$220%

These results show that increasing K generally improves performance, though a single-step perturbation ($K = 1$) slightly degrades the EER (1.02%), likely due to the well-known optimization instability of single-step adversarial training. As K increases, the EER gain becomes marginal beyond $K = 5$. Since Mel-BAP is a training-only regularizer, it incurs zero inference cost. However, during training, each PGD iteration requires an additional forward and backward pass through the Whisper-small encoder (which comprises the majority of the 252M total parameters). Therefore, higher K values substantially increase the computational burden (e.g., a $\sim$220% increase for $K = 10$). We selected $K = 5$ as the default setting, as it provides an optimal trade-off between robustness and training efficiency.

G Evaluation under Adaptive Adversarial Attacks

Table 11: Test-time adversarial evaluation on the BD-UT dataset. EER (%) is reported.

Attack Method	**Budget (ϵ)**	**Baseline EER (%)**	**Ours (Full) EER (%)**
No Attack (Clean)	-	1.19	**0.53**
PGD Attack (L_∞)	0.002	11.42	**3.81**
CW Attack (L_2)	2.0	15.85	**5.24**
Adaptive PGD (L_∞)	0.005	22.10	**7.93**

To comprehensively evaluate the true adversarial robustness of our framework and address the potential "blind spot" concern, we conducted test-time adversarial evaluations on the BD-UT dataset. This includes standard Projected Gradient Descent (PGD, L_∞), Carlini-Wagner (CW-L_2), and a specifically designed **Adaptive PGD** attack.

We simulated a strong adaptive white-box attacker who possesses full knowledge of our Mel-BAP prior (which restricts training-time perturbations to the 4–8 kHz band, denoted by the mask $\mathbf{M}_{4-8}$). This sophisticated adversary intentionally shifts all adversarial perturbation energy to the unprotected 0–4 kHz range ($\mathbf{M}_{0-4} = \mathbf{1} - \mathbf{M}_{4-8}$) to actively bypass our primary frequency defense:

$$\boldsymbol{\delta}_{adapt} = \arg \max_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}(\mathcal{F}_\theta(\mathbf{M} + \boldsymbol{\delta} \odot \mathbf{M}_{0-4}), y) \tag{9}$$

As summarized in Table 11, under the standard L_∞ PGD attack, the baseline Whisper-base model degrades substantially to 11.42% EER, whereas our full framework maintains a highly robust 3.81% EER. A similar trend is observed under the strong CW-L_2 attack (baseline 15.85% vs. ours 5.24%).

Most importantly, under the Adaptive PGD attack, the baseline further collapses to 22.10% EER. In sharp contrast, our method achieves 7.93% EER. This proves that our framework does not merely rely on trivial gradient masking in the target band, but learns a genuinely generalized, robust acoustic representation even when the attack is explicitly designed to exploit the defense mechanism. For all evaluations, we utilized identical attack protocols and budgets for both the baseline and our method to ensure a strictly fair comparison.

H Stratified Analysis of AAFI Routing Distributions

To examine whether AAFI routing varies across spoofing conditions, we conducted a stratified analysis of branch-utilization distributions across different attack types and severities. This analysis uses the Whisper-base Top-3 AAFI-only model, with Mel-BAP disabled to isolate the routing behavior.

As shown in Table 12, high-fidelity Neural Vocoder (NV) samples assign 59% of their selections to the wider 128/256-d branches, whereas standard TTS and noisy samples (SNR 10 dB) assign 64% and 48%, respectively, to the narrower 16/32-d branches. These aggregate distributions vary across input conditions and are consistent with condition-dependent multi-resolution selection. However, they provide associative rather than causal evidence and do not establish that artifact density directly determines the selected latent width.

Table 12: Stratified analysis of AAFI routing distributions across varying attack types and severities.

Category	**16/32-d**	**64-d**	**128/256-d**
NV (High-fidelity)	8%	33%	59%
TTS (Standard)	64%	22%	14%
Noisy (SNR 10dB)	48%	35%	17%

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims regarding SOTA performance (1.51% average EER) and the architectural efficiency of AAFI and Mel-BAP are directly supported by experimental results in Section 4 and Table 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Section 5 consolidates the principal limitations of this study, including sensitivity to routing capacity (Section 4.1), heterogeneous training protocols in cross-system comparisons (Section 4.3), the additional training cost of Mel-BAP (Appendix F), and unresolved real-time scalability in non-stationary acoustic environments. These limitations and trade-offs are discussed at their relevant locations and summarized explicitly in Section 5.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This study is primarily methodological and empirical; Information Bottleneck (IB) theory is used as a conceptual framework to motivate the architecture rather than for formal theoretical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide comprehensive details on the model architecture in Section 3 and disclose all training hyperparameters and dataset partitions in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The anonymous source code is provided in the supplemental material appended to this submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details are specified in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the high computational cost of training Whisper-based models, we report average EER across multiple benchmarks rather than per-run error bars, but we discuss result stability in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We disclose that all experiments were conducted using two NVIDIA A6000 GPUs in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research aims to improve digital trust and security by detecting audio deepfakes, strictly adhering to ethical standards.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the societal impact of combating misinformation and the potential dual-use concerns of deepfake detection in Sections 1 and 5.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our model is a defensive tool and does not provide generative capabilities, thus posing no risk for generating malicious deepfake content.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used datasets (e.g., ADD 2022, SpeechFake) and the Whisper backbone are properly cited and used according to their licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New modules like AAFI and Mel-BAP are fully documented in the methodology section and Appendix .4.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research did not involve crowdsourcing or experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research used publicly available datasets and did not require human subject approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for standard writing, editing, and formatting purposes, which does not require declaration under NeurIPS policy.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.