StableMark: Towards Codec-Robust Audio Watermarking via Attack Decomposition

Anonymous authors
Paper under double-blind review

Abstract

Audio watermarking is an emerging technology designed to embed imperceptible markers in audio signals for proactive deepfake detection and source tracing. However, these embedded markers are often susceptible to watermark removal attacks, particularly those leveraging neural codec resynthesis. While attack simulation during training has been widely adopted to enhance robustness against specific neural codec attacks, this approach often fails to generalize to unseen codecs or significantly degrades watermarked audio quality. To improve robustness against unseen neural codecs, we propose StableMark, a simulation-free audio watermarking framework. Our core idea is to first decompose nonlinear codec transformations into a set of attack-invariant functions that capture stable acoustic features preserved throughout resynthesis, and then construct the watermark detector as a set of linear combinations of these stable functions. In this paper, we start by proving that watermarking systems built upon attack-invariant functions are theoretically robust. Guided by this theory, we develop a method to derive generalizable stable functions through a combination of audio encoder pretraining and stable dimension selection. Extensive experiments across diverse neural codecs demonstrate that our framework significantly improves robustness against unseen neural codec attacks while maintaining high audio fidelity.

1 Introduction

The rapid advancement of high-fidelity audio generation models (Chen et al., 2025b; Majumder et al., 2024; Yuan et al., 2025a) has made it increasingly challenging for traditional audio deepfake detection (ADD) models to reliably distinguish deepfake audio from bona fide recordings (Dowerah et al., 2026). In response, audio watermarking emerges as a technology for proactive deepfake detection, source tracing and copyright protection. This technology embeds imperceptible markers into audio signals, which can be detected even after the audio is unintentionally transformed or maliciously edited.

Existing studies on audio watermarking (Pavlović et al., 2025; San Roman et al., 2024) generally focus on enhancing robustness against conventional audio edits (e.g., noise addition). However, most of them still remain vulnerable to deep attacks like neural codec resynthesis (Kovačević et al.). This allows an attacker to remove the watermark by copy-synthesizing the watermarked audio with a neural codec, yielding a clean version that remains perceptually indistinguishable from the original audio.

Current efforts on enhancing robustness against neural codec resynthesis attacks generally rely on attack simulation (San Roman et al., 2024; Chen et al., 2023). This paradigm inserts a differentiable approximation of the target neural codec between the watermark embedder and watermark detector during training. However, watermarking models trained with simulated attacks often suffer from two problems: **overfitting** and **quality degradation**. On the one hand, the resulting robustness is usually narrow, leaving the watermark vulnerable to unseen neural codecs. On the other hand, incorporating multiple or complex neural codecs into the training pipeline may introduce optimization instability, compromising the trade-off between robustness and watermarked audio quality.

To overcome the two limitations of attack simulation, this paper moves beyond codec-specific designs and instead establishes a system upon a set of attack-invariant functions shared among diverse

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

neural codecs. Specifically, our objective is to develop a watermarking system where the detector's output remains invariant under resynthesis by arbitrary neural codecs. We formally define this property as the "robustness" of audio watermarking models. To achieve this property, we first theoretically decompose neural codecs as a set of *eigenfunctions*. Each eigenfunction constitutes a transformation from a given audio signal into a real-valued scalar. For an eigenfunction with an eigenvalue of 1, this scalar remains constant even after the given audio is processed by a neural codec. Based on the assumption that there exists a non-empty set of eigenfunctions shared among different neural codecs, we derive a sufficient robustness condition: *A watermark system is robust if the watermark detector is a linear combination of eigenfunctions with eigenvalues equal to 1.*

This theoretical finding raises a critical practical question: how can we find a set of eigenfunctions shared among different neural codecs? Our solution is to model these eigenfunctions with a neural audio encoder, where each output dimension is optimized to represent an eigenfunction. Furthermore, to ensure that attack-invariant functions learnt on seen codecs generalize to unseen ones, we perform stable dimension selection by identifying features that remain stable against a development set of codecs. The final watermark detector is constructed as a linear combination of these selected stable functions, with the watermark embedded through adversarial optimization. Experiments show that the proposed method significantly improves watermark detection accuracy under both seen and unseen neural codec attacks, while maintaining high audio quality, high capacity, and robustness against other conventional attacks.

The contribution of this paper is three-fold: 1) **Theoretically**, we derive the robustness condition of robust audio watermarking; 2) **Methodologically**, we design StableMark, an audio watermarking framework based on generalizable attack-invariant functions and proposed a method to derive such functions with audio encoder pretraining and stable dimension selection; 3) **Empirically**, extensive experiments confirm that our method achieves superior accuracy under both seen and unseen neural codec attacks while maintaining imperceptibility.

2 Related Works

Audio watermarking is an information hiding technique that encodes secret messages, either zero-bit labels (Juvela & Wang, 2025; Yuan et al., 2025b) or multi-bit binary strings (San Roman et al., 2024; Liu et al., 2023b; Pavlović et al., 2025), into audio waveforms. An audio watermarking model must balance three conflicting requirements: 1) Imperceptibility: ensuring the watermarked audio is audibly similar to the original; 2) High capacity: allowing sufficient message bits within a unit-length window; 3) Robustness: the watermark message should be resilient to various transformations and attacks. In this work, we focus on enhancing robustness of multi-bit audio watermarking systems.

Traditional audio watermarking methods (Boney et al., 1996; Yeo & Kim, 2003) focus on robustness against transmission distortions like filtering, quantization and conventional compression. However, since they rely on hand-crafted rules, these methods are still susceptible to strong attacks (Chen et al., 2023). The application of neural networks significantly simplifies the design of audio watermarking models, enhancing robustness, imperceptibility and capacity simultaneously (Pavlović et al., 2022). Following works explored approaches like attack simulation (Liu et al., 2023a;b), architectural modifications (San Roman et al., 2024; Pavlović et al., 2025) and model ensembling (Xu et al., 2025; Yuan et al., 2025b) to defend more complex attacks, such as replay (Liu et al., 2023a) and cropping (Pavlović et al., 2025). However, these paradigms still face inherent limitations: Attack simulation often triggers overfitting or optimization instability, while architectural modifications generally focus merely on specific attacks. Furthermore, ensembling strategies are, in essence, an ensemble of single watermarking systems. Therefore, their performance may still be limited by the robustness of the base watermarking model.

Besides the aforementioned conventional attacks, most existing audio watermarking methods remain susceptible to deep attacks like neural codec resynthesis (Özer et al., 2025). To tackle this, DiscreteWM (Ji et al., 2025b) attempted to embed watermark in discrete latent space of a Vector-Quantized Variational AutoEncoder (VQ-VAE). However, accuracy on neural codecs is still not reported. Other works (Juvela & Wang, 2025; Yuan et al., 2025b) tried to develop zero-bit audio watermarking systems robust to neural codecs. In contrast, our work goes beyond single labels and head towards codec-robust multi-bit audio watermarking, which benefits not only proactive deepfake detection but also source tracing and copyright protection.

3 THEORETICALLY ROBUST AUDIO WATERMARKING

The goal of this section is to construct a watermarking system that is theoretically robust to an arbitrary attack $\mathcal{A}$. We begin by formally defining "robustness" of an audio watermarking system in Section 3.1. Then in Section 3.2, we decompose the attack $\mathcal{A}$ to obtain a set of eigenfunctions, and derive sufficient conditions to achieve robustness via attack decomposition. Guided by these robustness conditions, we construct a theoretically robust audio watermarking system where the watermark decoder is a linear combination of eigenfunctions with eigenvalues equal to 1.

3.1 PROBLEM FORMULATION

Let $\mathcal{X} \subseteq \mathbb{R}^T$ be a fixed-length natural audio waveform space and $\mathbf{m} \in \{0,1\}^N$ be an arbitrary N-bit message. Consider an audio watermarking system with embedder $\mathcal{G} : \mathcal{X} \times \{0,1\}^N \to \mathcal{X}$ and detector $\mathcal{D} : \mathcal{X} \to \{0,1\}^N$. We decompose $\mathcal{D}$ into N bitwise real-valued scoring functions $\{f_i\}_{i=1}^N$ where we assume $f_i(\mathbf{x}) \neq 0$. Denote the watermarked audio as $\mathbf{x}_w = \mathcal{G}(\mathbf{x}, \mathbf{m})$. Each message bit $\hat{m}_i$ is predicted by $\hat{m}_i = \mathbb{I}[f_i(\mathbf{x}_w) > 0]$.

Definition 3.1 (Robust Audio Watermarking)**.** A watermarking system $(\mathcal{G}, \mathcal{D})$ is considered robust to deterministic attack $\mathcal{A} : \mathcal{X} \to \mathcal{X}$ if $\forall \mathbf{x} \in \mathcal{X}$, $\mathbf{m} \in \{0,1\}^N$ and $i \in \{1, \cdots, N\}$, we have:

$$\mathbb{I}[f_i(\mathbf{x}_w) > 0] = \mathbb{I}[f_i(\mathcal{A}(\mathbf{x}_w)) > 0] = m_i \tag{1}$$

3.2 ROBUSTNESS CONDITIONS

Below we derive sufficient robustness conditions via attack decomposition. While direct spectral decomposition is intractable for non-linear operator $\mathcal{A}$, we introduce a "Koopman operator" $\mathcal{K}_\mathcal{A}$ as an analytical proxy to lift $\mathcal{A}$ into a linear function space for decomposition (Mauroy et al., 2020).

Definition 3.2 (Koopman Operator for Attacks)**.** Let $\mathcal{F}$ be the space of observable functions $\psi : \mathcal{X} \to \mathbb{R}$. The Koopman operator $\mathcal{K}_\mathcal{A} : \mathcal{F} \to \mathcal{F}$ associated with attack $\mathcal{A}$ is defined as:

$$(\mathcal{K}_\mathcal{A}\psi)(\mathbf{x}) := \psi(\mathcal{A}(\mathbf{x})), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{2}$$

Note that unlike attack operator $\mathcal{A}$, the operator $\mathcal{K}_\mathcal{A}$ is linear. We can thus perform spectral analysis on $\mathcal{K}_\mathcal{A}$. Our first step is to define *eigenfunctions*, a class of functions that behave linearly under non-linear attacks.

Definition 3.3 (Eigenfunctions of a Koopman Operator)**.** A function $\psi_k \in \mathcal{F}$ is an eigenfunction of $\mathcal{K}_\mathcal{A}$ with eigenvalue λ_k if

$$\mathcal{K}_\mathcal{A}\psi_k = \lambda_k \psi_k, \tag{3}$$

or equivalently,

$$\psi_k(\mathcal{A}(\mathbf{x})) = \lambda_k \psi_k(\mathbf{x}). \tag{4}$$

Assume that we have identified a set of linearly independent eigenfunctions $\{\psi_k\}_{k=1}^r$, we then consider constructing each bitwise watermark detector $f_i : \mathcal{X} \to \mathbb{R}$ within the subspace they span, so that f_i can inherit these eigenfunctions' stability under attack $\mathcal{A}$:

$$f_i(\mathbf{x}_w) = \sum_{k=1}^{r} w_{ik} \psi_k(\mathbf{x}_w), \tag{5}$$

where $w_{ik} \in \mathbb{R}$ are the spectral weights of the detector. Applying the attack $\mathcal{A}$ to the audio $\mathbf{x}_w$ is equivalent to modulating the detector's components by their eigenvalues:

$$f_i(\mathcal{A}(\mathbf{x}_w)) = \sum_{k=1}^{r} \lambda_k w_{ik} \psi_k(\mathbf{x}_w). \tag{6}$$

The following part shows that f_i constructed as in Equation (5) is robust against attack $\mathcal{A}$ as long as it satisfies certain conditions. Substituting Equations (5) and (6) into Definition 3.1, we first derive the equivalent robustness condition:

Theorem 3.4 (Equivalent Robustness Condition)**.** *Let $\mathbf{m}' = 2\mathbf{m} - \mathbf{1}$ be the watermark message mapped from $\{0,1\}^N$ to $\{-1,1\}^N$. The detector is robust at $\mathbf{x}_w$ if and only if:*

$$\begin{cases} m_i' \sum_{k=1}^{r} w_{ik}\psi_k(\mathbf{x}_w) & > 0 \\ m_i' \sum_{k=1}^{r} \lambda_k w_{ik}\psi_k(\mathbf{x}_w) & > 0 \end{cases} \tag{7}$$

To derive a more interpretable condition on the weights $\{w_{ik}\}_{k=1}^{r}$, we decompose the spectrum into robust and fragile sets based on the eigenvalues. Let $I_{\|} = \{k \mid \lambda_k = 1\}$ and $I_{\perp} = \{k \mid \lambda_k \neq 1\}$. We split the attacked response into a *robust term* $f_{i,\|}(\mathbf{x}_w)$ and a *fragile term* $f_{i,\perp}(\mathbf{x}_w)$:

$$f_i(\mathcal{A}(\mathbf{x}_w)) = \underbrace{\sum_{k\in I_{\|}} w_{ik}\psi_k(\mathbf{x}_w)}_{f_{i,\|}(\mathbf{x}_w)} + \underbrace{\sum_{k\in I_{\perp}} \lambda_k w_{ik}\psi_k(\mathbf{x}_w)}_{f_{i,\perp}(\mathbf{x}_w)}. \tag{8}$$

Then the last line of Equation (7) can be reformulated as:

$$m_i' f_{i,\|}(\mathbf{x}_w) > -m_i' f_{i,\perp}(\mathbf{x}_w). \tag{9}$$

Since the sign and magnitude of the fragile residual $f_{i,\perp}(\mathbf{x})$ are unpredictable, we derive sufficient conditions that guarantee robustness against the worst-case interference from the fragile components.

Theorem 3.5 (Weak Robustness Condition)**.** *The watermark system is robust at $\mathbf{x}_w$ if 1) The original detection is correct; 2) The margin of the robust component exceeds the magnitude of the fragile component. Formally,*

$$\begin{cases} m_i' f_i(\mathbf{x}_w) > 0 \\ m_i' f_{i,\|}(\mathbf{x}_w) > |f_{i,\perp}(\mathbf{x}_w)| \end{cases} \tag{10}$$

Theorem 3.6 (Strong Robustness Condition)**.** *The watermark system is robust against attack $\mathcal{A}$ if: 1) The original detection is correct; 2) The detector is a linear combination of eigenfunctions with $\lambda_k = 1$. Formally,*

$$\begin{cases} m_i' \sum_{k\in I_{||}} w_{ik}\psi_k(x_w) > 0 \quad \forall k \in I_{\|} \\ w_{ik} = 0 \quad \forall k \in I_{\perp}. \end{cases} \tag{11}$$

Theorem 3.6 indicates that as long as a watermark detector is a linear combination of eigenfunctions ψ_k with eigenvalues $\lambda_k = 1$, it is provably robust under attack $\mathcal{A}$. In other words, *robust watermark detection can be guaranteed as long as we identify a set of eigenfunctions with eigenvalues of 1*. For brevity, we refer to eigenfunctions ψ_k with eigenvalues $\lambda_k = 1$ as *robust eigenfunctions*.

4 Implementation

Guided by Theorems 3.5 and 3.6, this section introduces a bottom-up approach to build a robust audio watermarking system by first approximating a set of robust eigenfunctions via audio encoder pretraining and dimension selection (see Section 4.1), and then building an audio watermarking system based on these approximated robust eigenfunctions (see Section 4.2).

4.1 Approximation of Robust Eigenfunctions

In order to approximate robust eigenfunctions defined in Section 3, we design a two-stage process consisting of a **learning phase** and a **selection phase**. In the learning phase, an audio encoder learns representations with reduced sensitivity to codec resynthesis. Then in the selection phase, r most stable dimensions are selected from the representation space as approximated stable eigenfunctions.

The architecture of the audio encoder is depicted in Figure 1 (a). The network works on the mel-spectrogram domain and is composed of 3 convolutional blocks and a linear output layer. In each convolutional block, we stack a single-strided 1D convolutional layer with kernel size of 1, a leaky-ReLU activation function, and an instance normalization layer. Note that all these operations are performed independently on every frame of length T. Finally, the linear output layer maps the feature of each frame into a latent vector.

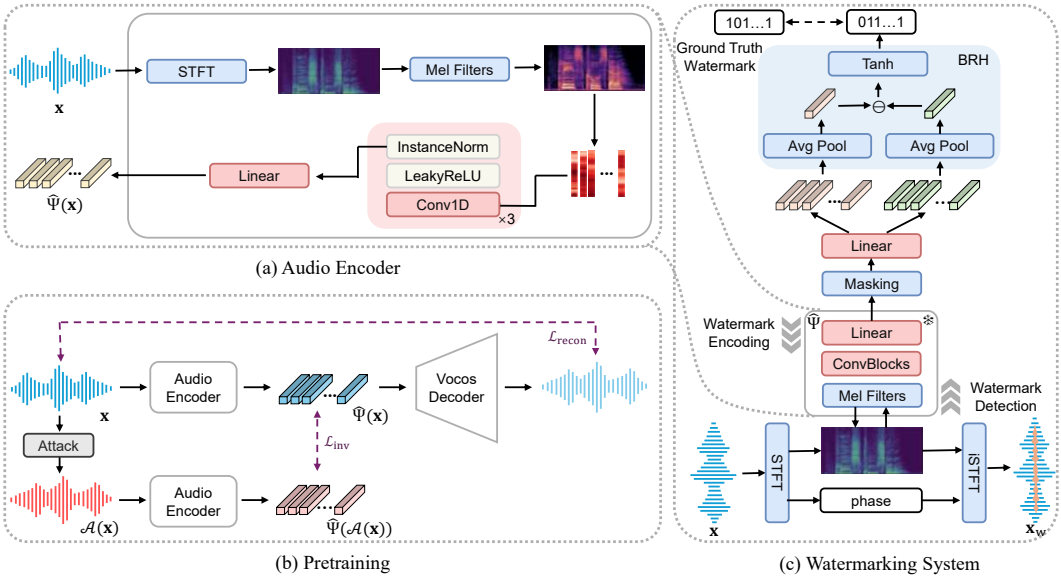

Figure 1: (a) Network architecture of the audio encoder, whose core module is a frame-by-frame CNN working on Mel spectrograms. (b) Pretraining pipeline of the audio encoder. We augment the conventional autoencoder pipeline with an invariance loss to align the latent features extracted from original and attacked audio. (c) Embedding and detection process of audio watermarking system. Watermark is embedded by iteratively optimizing STFT amplitude spectrogram via gradient descent.

The audio encoder is trained in an autoencoder framework with a Vocos decoder (Siuzdak), a multi-period discriminator (MPD) and a multi-resolution discriminator (MRD). The training objectives are designed to fulfill two theoretical requirements: 1) **Invariance**: Each latent dimension should approximate a robust eigenfunction. Specifically, Equation (4) with $\lambda_k = 1$ indicates that $\mathbb{E}_{\mathbf{x}}[|\hat{\psi}_k(\mathcal{A}(\mathbf{x})) - \hat{\psi}_k(\mathbf{x})|]$ should be minimized for each output dimension $\hat{\psi}_k$; 2) **Diversity**: The approximated eigenfunctions should be sufficiently diverse for spanning a non-trivial (high rank) subspace that can represent various detectors for multi-bit watermark embedding.

To satisfy the invariance requirement, we apply an invariance loss $\mathcal{L}_{inv}$ in the latent space: During training, both the original audio $\mathbf{x}$ and the attacked audio $\mathcal{A}(\mathbf{x})$ are given as input of the audio encoder, yielding clean latent features $\hat{\Psi}(\mathbf{x}) = [\,\hat{\psi}_1(\mathbf{x})\ \hat{\psi}_2(\mathbf{x})\ \cdots\ \hat{\psi}_d(\mathbf{x})\,]^\top$ and attacked latent features $\hat{\Psi}(\mathcal{A}(\mathbf{x})) = [\,\hat{\psi}_1(\mathcal{A}(\mathbf{x}))\ \hat{\psi}_2(\mathcal{A}(\mathbf{x}))\ \cdots\ \hat{\psi}_d(\mathcal{A}(\mathbf{x}))\,]^\top$ respectively. A direct approach to encourage invariance is to minimize the ℓ_2 distance. However, as indicated in Appendix D.1, strictly optimizing towards theoretically robust eigenfunctions significantly restricts the space for watermark embedding, leading to limited detection accuracy. Therefore, we instead adopt cosine similarity loss to find an approximation set of robust eigenfunctions with $\lambda \simeq 1$ that may satisfy Theorem 3.5:

$$\mathcal{L}_{inv} = -\mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\left[\frac{\hat{\Psi}(\mathbf{x}) \cdot \hat{\Psi}(\mathcal{A}(\mathbf{x}))}{\|\hat{\Psi}(\mathbf{x})\|_2 \|\hat{\Psi}(\mathcal{A}(\mathbf{x}))\|_2}\right] \tag{12}$$

The diversity requirement can be satisfied by the autoencoder reconstruction loss, since high-quality reconstruction audio can be produced only when the audio encoder encodes sufficient acoustic features from the original audio. Specifically, the reconstruction loss is a weighted sum of the GAN loss $\mathcal{L}_D$ and the ℓ_1 distance $\mathcal{L}_{mel}$ between Mel spectrogram of original and reconstructed audio:

$$\mathcal{L}_{recon} = \lambda_D \mathcal{L}_D + \lambda_{mel} \mathcal{L}_{mel}. \tag{13}$$

Compared to exactly equivalence constraints, learning with cosine similarity loss yields more functions with varying levels of stability. We therefore introduce a subsequent selection stage to remove the functions with $\lambda >> 1$ or $\lambda << 1$. Specifically, the selection stage identifies a subset of $r < d$ dimensions that mostly resembles the set of robust eigenfunctions $I_{\|}$ required in Theorem 3.6. Specifically, we propose Universal Stability Index (USI) to quantify stability of each dimension. USI aggregates four statistical metrics derived from the Wasserstein-1 (W_1) distance between distributions of each dimension of clean and codec-resynthesized audio features: 1) **Invariance Score** S_{inv}: The mean W_1 distance between clean features and those perturbed by both seen and unseen

codecs, representing overall stability; 2) **Transferability Score** S_{trans}: The W_1 distance between features processed by seen and unseen codecs, measuring generalization ability from seen to unseen; 3) **Cohesion Score** S_{coh}: The average pairwise W_1 distance among all unseen codecs, capturing stability among unseen; 4) **Consistency Score** S_{std}: Standard deviation of W_1 distances across unseen codecs, penalizing dimensions with inconsistent stability across diverse architectures. Final USI_k for k-th dimension is a weighted sum of Z-scores of these four metrics:

$$USI_k = \alpha Z(S_{inv,k}) + \beta Z(S_{trans,k}) + \gamma Z(S_{coh,k}) + \delta Z(S_{std,k}) \tag{14}$$

where $\alpha, \beta, \gamma, \delta$ are hyperparameters controlling the selection bias toward different stability facets.

Finally, we rank all d dimensions in ascending order of USI_k and identify the top r most stable dimensions as stable eigenfunctions. In the watermarking system, this selection mechanism is implemented as a masking layer. Specifically, we define a binary mask $\mathbf{M} \in \{0,1\}^d$, where $M_k = 1$ only for $k \in I_{\|}$. This mask is element-wise multiplied with the encoder's output $\hat{\Psi}(\mathbf{x})$ to zero out unstable dimensions, ensuring that only the robust subspace $I_{\|}$ contributes to the final bitwise prediction. In practice, we treat codecs in the development set as unseen for estimating USI score.

4.2 Eigenfunction-Based Watermarking System

Based on the approximated stable eigenfunctions, we build a detector-only robust watermarking system. Guided by Equation (5), the watermark detector is constructed by appending a randomly-initialized linear layer and a bitwise readout head (BRH) after the pretrained audio encoder. As shown in Figure 1 (c), for each watermark bit and each frame, the linear layer predicts a score for "1" and a score for "0". Then we average the two scores across all frames, take the difference between the "1" and "0" score, and map it into $(-1, 1)$ with a tanh function. The predicted score of the i-th bit is

$$\hat{m_i} = \tanh(\sum_{k=1}^{r} w_{ik}\hat{\psi}_k(\mathbf{x})). \tag{15}$$

Since each selected output dimension $\hat{\psi}_k$ satisfies $\hat{\psi}_k(\mathcal{A}(\mathbf{x})) \approx \hat{\psi}_k(\mathbf{x})$, the predicted score $\hat{m_i}$ remains invariant to attacks applied to the input $\mathbf{x}$.

Inspired by the embedder-free design of AWARE (Pavlović et al., 2025), our system embeds watermark by iteratively optimizing the input audio to minimize a watermark detection loss $\mathcal{L}_{wm}$, so that the decoded watermark $\hat{\mathbf{m}}$ could be close to the ground-truth watermark $\mathbf{m}^*$:

$$\mathcal{L}_{wm} = \frac{1}{N}\|\mathbf{m}^* - \hat{\mathbf{m}}\|_2^2 - \lambda\frac{1}{N}\sum_{i=1}^{N}|\hat{m_i}|. \tag{16}$$

Following (Pavlović et al., 2025), we perform watermark embedding and detection in the STFT amplitude domain. Specifically, we embed the watermark by iteratively updating the STFT amplitude spectrogram via gradient backpropagation to minimize $\mathcal{L}_{wm}$. Meanwhile, the signal-perturbation ratio $\frac{|\Delta|}{|STFT(\mathbf{x})|} = |\frac{|STFT(\mathbf{x}_w)|}{|STFT(\mathbf{x})|} - 1|$ is limited under a threshold $\eta = 10^{-\tau_{dB}/20}$ where τ_{dB} is a hyperparameter. The modified amplitude is then combined with the original phase to reconstruct the watermarked signal via inverse STFT. Throughout the embedding process, only the amplitude spectrogram is updated, leaving the pretrained audio encoder and the linear layer frozen. During watermark detection, a lightweight WebRTC Voice Activity Detector (VAD) module is first applied to ensure that watermark is detected on only voiced segments. The audio encoder then extracts frame-wise latent features from the STFT amplitude spectrogram. Subsequently, the linear layer and the BRH aggregate these attack-invariant features and produce the final bitwise prediction: a positive output indicates bit "1", and a negative output indicates bit "0".

5 Experiments

5.1 Experimental Setup

We train the audio encoder on the CoRS subset of the CodecFake+ dataset (Chen et al., 2025a). This dataset contains VCTK (Yamagishi et al., 2019) speech copy-synthesized with 31 neural codecs covering 21 types of model architectures (details in Table 6 of Appendix). To ensure a fair comparison with baselines that have seen EnCodec-6b24k (Défossez et al., 2022) during training, our training

Table 1: Bitwise/existence Accuracy (%) against neural codec attacks and AI attacks. Baselines with '*' includes EnCodec-6b24k as a simulated attack (SA) during training.

Model	nbits	Clean Acc.	Neural Codec Attacks			AI Attacks
			Train	Dev	Test	
AudioSeal*	16	**100.0**/99.58	63.48/51.24	60.16/61.43	47.60/50.28	58.79/58.98
WavMark	16	**100.0**/99.62	53.83/50.75	58.05/57.73	53.82/50.21	57.82/51.65
Timbre	30	**100.0/100.00**	63.22/54.00	65.81/68.26	60.77/55.54	83.00/84.01
Timbre-SA*	30	99.97/99.04	**90.26/99.79**	72.58/75.16	66.66/71.51	87.79/**90.55**
AWARE	30	98.30/99.41	69.42/82.82	68.69/73.74	61.49/65.80	86.81/87.56
StableMark (ours)	30	99.86/99.50	82.00/96.06	**74.03/76.56**	**69.05/76.66**	**88.14**/88.87
w/o pretraining	30	98.85/99.94	56.46/50.00	59.75/60.78	51.16/50.67	77.63/83.92
w/o inv. loss	30	99.66/99.96	74.32/87.19	70.12/74.40	62.31/65.50	84.65/87.17
w/o recon. loss	30	62.07/75.12	51.75/52.36	52.09/54.08	48.13/46.38	59.30/71.96
w/o dim. selection	30	99.24/98.10	76.48/94.41	69.77/74.77	64.28/70.25	84.96/87.86
w/o VAD	30	99.92/99.98	77.72/92.72	72.68/75.34	66.18/68.60	86.50/88.15

set is restricted to audio from EnCodec-6b24k and its bona fide counterparts unless otherwise stated. From the remaining codecs, we randomly assign 6 to the development set for dimension selection, while the other 22 are reserved for testing. All test samples are from the LibriSpeech test-clean corpus. For brevity, each codec is referenced by an alphabetical ID in the following part. Detailed mapping of these IDs to codec names and their respective partitions can be found in Appendix C.1.

We evaluate each watermarking system on two dimensions: robustness and imperceptibility. Robustness is assessed through bitwise accuracy of the payload bitstring and existence accuracy of the binary watermarked/unwatermarked label. To calculate existence accuracy, the watermark detector processes both a watermarked and an unwatermarked version of the test set. A sample is classified as "watermarked" if its bitwise accuracy exceeds a predefined threshold τ, which is determined as in Appendix C.5. Furthermore, to ensure the embedded watermark remains imperceptible, we report objective audio quality metrics including PESQ (Rix et al., 2002), STOI (Taal et al., 2011), VISQOL(Sloan et al., 2017) and NISQA(Mittag et al., 2021), and a subjective metric mean opinion score (MOS), on quality of watermarked audio. Specifically, MOS is obtained through blind listening tests where participants rate the quality of 10 watermarked audios per system on a 5-point scale: 5 represents perfect imperceptibility, while 1 represents obvious distortion.

We compare our method against state-of-the-art multi-bit audio watermarking systems including AudioSeal (San Roman et al., 2024), WavMark (Chen et al., 2023), Timbre (Liu et al., 2023b) and AWARE (Pavlović et al., 2025). Among these baselines, only AudioSeal includes EnCodec-6b24k as a simulated attack (SA) during training. Additionally, we implement Timbre-SA, an enhanced version of Timbre trained with the same EnCodec-6b24k simulated attack. This allows us to directly compare the effectiveness of our decomposition-based approach against the standard attack simulation strategy.

5.2 Results

5.2.1 Robustness

Table 1 demonstrates the robustness of our proposed system against two groups of watermark removal attacks: neural-codec resynthesis attacks and other types of AI attacks (neural vocoder, VAE, diffusion model resynthesis and speech enhancement). In the "Clean Acc." column, we also report watermark prediction accuracy when no attack is applied to the watermarked audio. Without attacks, both the baselines and our system achieve nearly 100% bitwise and existence accuracy, indicating utility of the proposed method under the clean setting.

For robustness against neural codec resynthesis attacks, accuracy on seen (Train & Dev) and Unseen (Test) codecs are both reported. Results show that while most baselines fail to generalize to unseen codecs, our method achieves superior bitwise accuracy on these cases while remaining competitive on seen ones. Notably, although Timbre-SA shows competitive robustness, it significantly degrades watermarked audio quality as indicated in Table 2. The left part of Figure 2 further compares the per-attack bit accuracy of our method against several baselines. Except Timbre-SA, the best performance baseline achieves over 60% bit accuracy merely on half of unseen test codecs, whereas our method achieves over 60% bit accuracy on 17 out of 22 test codecs.

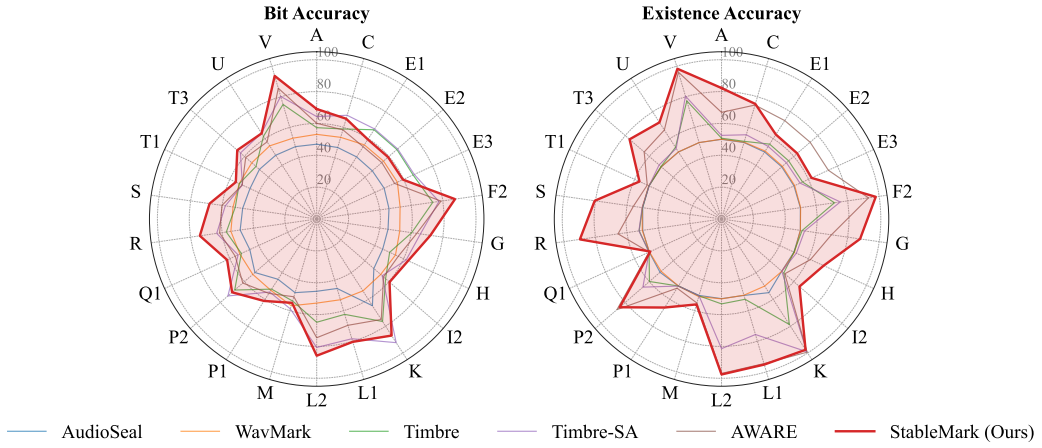

Figure 2: Bitwise and existence accuracy on each unseen codec.

Table 2: Comparison of watermarked audio quality.

Model	PESQ↑	STOI↑	VISQOL↑	NISQA↑	MOS↑
AudioSeal (San Roman et al., 2024)	4.422	0.998	4.305	4.006	4.391
WavMark (Chen et al., 2023)	4.187	0.997	3.686	4.054	4.000
Timbre (Liu et al., 2023b)	3.856	0.990	3.606	3.665	4.000
Timbre-SA	1.898	0.884	2.267	2.399	1.406
AWARE (Pavlović et al., 2025)	3.911	0.975	3.809	3.572	3.953
StableMark (ours)	3.950	0.972	3.694	3.573	4.141
w/o pretrain	3.908	0.975	3.813	3.567	3.922
w/o inv. loss	3.937	0.972	3.704	3.599	3.875
w/o recon loss	3.959	0.966	3.706	3.607	3.906
w/o dim. selection	3.910	0.972	3.686	3.647	3.984
w/o VAD	3.910	0.972	3.676	3.635	4.000

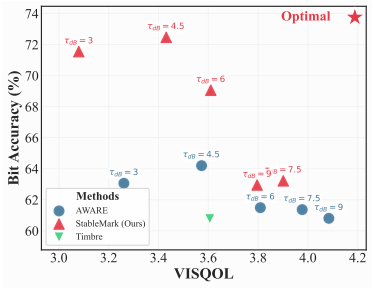

Figure 3: Trade-off between bitwise accuracy and audio quality.

The advantage of our method is amplified on existence accuracy. As shown in Table 1, our method achieves an existence accuracy of over 75%, significantly surpassing all baselines. Furthermore, the right part of Figure 2 shows that compared with baselines, our method achieves the highest existence accuracy on over 70% of unseen codecs.

Additionally, we find that eigenfunctions learnt on seen codecs can generalize to not only unseen codecs but also other types of AI attacks. This is supported by the superior bit accuracy under AI attacks as reported in the last column of Table 1. Detailed accuracy under each AI attack, as well as accuracy under conventional audio edits can be found in Appendix D.3 and D.4. Performance with higher watermark capacity is also discussed in Appendix D.5.

5.2.2 Imperceptibility

Table 2 summarizes watermarked audio quality of different systems. While some baselines like AudioSeal and WavMark achieve higher imperceptibility, their robustness is significantly lower than ours as indicated in Table 1. Except AudioSeal, our method ranks first in terms of MOS, whereas our subjective quality scores remain competitive against baselines. Since subject metrics may mistake watermark for noise during evaluation, we can conclude that audio watermarked by our system is perceptually indistinguishable from the original in most scenarios.

Additionally, Figure 3 illustrates the trade-off between robustness and imperceptibility as the embedding threshold τ_{dB} varies. Compared with AWARE and Timbre baselines, our method occupies a position closer to the upper-right region, representing a better robustness-imperceptibility trade-off.

5.3 Ablation Studies

In lower parts of Tables 1 and 2, we conduct ablation studies to verify the effectiveness of key components in the proposed method: 1) **Audio Encoder Pretraining**, without which both bit and existence accuracy collapse to approximately 50%. This suggests that randomly-initialized neural network is

Table 3: Bit accuracy on codecs with unseen architectures when using different codecs for training.

Seen Codec	A	C	E2	G	H	I2	J	M	P2	Q1	R	S	T1	U	W	MEAN	# Params.
Spectral Codecs (V)	61.61	58.56	53.63	64.43	55.45	54.68	46.33	47.91	57.13	55.52	64.75	60.53	50.41	56.23	53.18	56.02	64.33M
DAC (B)	59.86	57.31	52.9	63.31	54.77	53.38	47.19	49.25	60.99	54.61	64.6	59.6	57.89	55.93	52.80	56.29	74.71M
Vocos_encodec_12 (L1)	62.23	60.26	53.77	65.35	56.42	56.32	47.33	49.02	59.91	56.28	66.46	61.06	50.97	59.02	**56.37**	57.38	24.93M
EnCodec-6b24k (D1)	64.21	60.73	55.62	67.26	58.88	56.85	46.97	51.08	64.64	58.39	67.88	62.94	51.43	59.93	44.68	58.10	14.85M
FunCodec (F1)	66.86	64.51	59.80	69.59	61.38	60.79	54.45	56.8	63.35	61.49	70.51	65.86	56.58	64.24	52.50	61.91	4.73M
HILCodec (K)	**69.59**	**66.22**	57.48	**72.07**	63.24	59.95	51.26	50.35	64.04	62.36	**74.46**	**68.10**	51.37	**66.02**	55.65	62.14	15.66M
EnCodec-12b24k (D2)	66.65	65.00	**64.45**	67.83	**64.08**	**62.32**	**59.38**	**60.08**	**67.91**	**63.36**	70.36	65.95	**59.18**	64.16	48.10	**63.25**	14.85M

not robust against neural codec attacks. 2) **Invariance Loss** and **Reconstruction Loss**: Removing either $\mathcal{L}_{inv}$ or $\mathcal{L}_{recon}$ would result in a drop in accuracy under codec attacks, indicating their crucial roles in encouraging the audio encoder to approximate robust eigenfunctions. 3) **Dimension Selection**, without which robustness would drop due to noise introduced by unstable dimensions. 4) **VAD**, which slightly improves robustness by restricting watermark on voiced segments.

5.4 Analysis

In this section, we discuss two design choices and their influence on performance: number of eigenfunctions r and choice of codecs for audio encoder pretraining.

Number of Eigenfunctions The number of selected dimensions r controls the trade-off between capacity and robustness. We determine the optimal r by analyzing performance of the proposed method on the development set. As illustrated in Figure 4, bit accuracy on development set first rises as the dimension rank increases and reaches a peak plateau at $r = 46$. Then it starts to decline as more dimensions are included. To explain this phenomenon, we rank all $d = 128$ dimensions in increasing order of USI and plot out USI of each dimension in Symmetric Logarithm (Symlog) scale. When r is too small, the selected "stable subspace" will not be large enough to encode a 30-bit binary message. On the contrary, when r becomes large, accuracy under neural codec attacks would drop due to reliance on unstable dimensions with high USIs. Our choice of r can be further validated by the optimal bit accuracy on testing set, indicating that stable dimensions selected on development codecs can generalize to unseen test codecs.

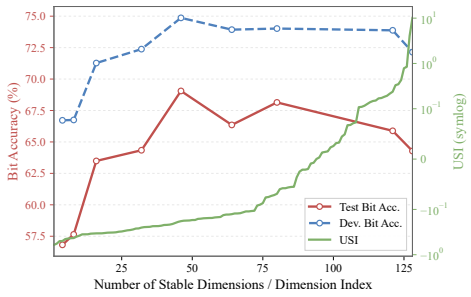

Figure 4: Left: Bitwise accuracy of our system when using different number of stable dimensions r; Right: USI of each dimension in symlog scale.

Types of Seen Codecs In the main experiment, we utilize EnCodec-6b24k as the seen codec for fair comparison with baselines. In this section, we demonstrate that switching to other types of training codecs like EnCodec-12b24k can further improve accuracy on unseen codec attacks. To isolate pretraining effects, we compare robustness of systems trained with different seen codecs without conducting dimension selection on the development set. Additionally, we employ a dedicated test set different from that in Table 6 to exclude all candidate training codecs during evaluation. As shown in Table 3, the choice of codec for training significantly influences the system's robustness and generalization ability. Specifically, we find that using light-weight codecs with fewer parameters as training codecs generally improves robustness of the proposed system. We suppose that this is because lighter codecs introduce fewer synthesis artifacts unique to the training codec. Therefore, eigenfunctions learnt on these codecs generalize better to unseen ones.

6 Conclusion

In this paper, we proposed StableMark, a novel framework for codec-robust audio watermarking based on attack decomposition. By decomposing codec attacks as a series of eigenfunctions, we derive theoretical conditions for robust audio watermarking, train an audio encoder to model stable eigenfunctions, and design an audio watermarking system based on the learned eigenfunctions. Experimental results on 22 test codecs demonstrate that our method effectively improves robustness against unseen neural codec while maintaining imperceptibility. By shifting the focus from attack simulation to attack decomposition with eigenfunctions shared among diverse neural codecs, our work provides a new perspective on defending the challenging neural codec resynthesis attack against audio watermarking systems.

AI USE STATEMENT

In this work, we used generative AI tools for coding and polishing writing. We have not used generative AI tools for retrieval and discovery, research ideation or execution, generating synthetic datasets, or proving mathematical claims. We have reviewed all AI-assisted work. We checked LLM-refined texts, and LLM-generated code was verified and tested for correctness. We take responsibility for final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper contains a MOS evaluation of watermarked audio quality. The MOS study involved 8 student participants, who received no compensation. Each participant rated 8 sets of watermarked audio samples from 11 anonymized systems on a 1–5 scale in a blind listening test. The test was reference-based: listeners had access to the original unwatermarked recording. The watermarked samples were randomly shuffled before being presented to the listeners.

REPRODUCIBILITY STATEMENT

We plan to open source our codes upon acceptance. To enhance reproducibility, we also provide implementation details in Appendix C.

REFERENCES

Sunghwan Ahn, Beom Jun Woo, Min Hyun Han, Chanyeong Moon, and Nam Soo Kim. Hilcodec: High-fidelity and lightweight neural audio codec. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1517–1530, 2024.

L. Boney, A.H. Tewfik, and K.N. Hamdy. Digital watermarks for audio signals. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, pp. 473–480, 1996. doi: 10.1109/MMCS.1996.535015.

Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

Xuanjun Chen, Jiawei Du, Haibin Wu, Lin Zhang, I Lin, I Chiu, Wenze Ren, Yuan Tseng, Yu Tsao, Jyh-Shing Roger Jang, et al. Codecfake+: A large-scale neural audio codec-based deepfake speech dataset. *arXiv preprint arXiv:2501.08238*, 2025a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025b.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2022.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Mathew Magimai.-Doss. Speech df arena: A leaderboard for speech deepfake detection models. *IEEE Open Journal of Signal Processing*, pp. 1–9, 2026. doi: 10.1109/OJSP.2026.3652496.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 591–595. IEEE, 2024.

Haohan Guo, Fenglong Xie, Kun Xie, Dongchao Yang, Dake Guo, Xixin Wu, and Helen Meng. Socodec: A semantic-ordered multi-stream speech codec for efficient language model based text-to-speech synthesis. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 645–651. IEEE, 2024.

Shengpeng Ji, Minghui Fang, Ziyue Jiang, Siqi Zheng, Qian Chen, Rongjie Huang, Jialung Zuo, Shulei Wang, and Zhou Zhao. Language-codec: Reducing the gaps between discrete codec representation and speech language models. *arXiv e-prints*, pp. arXiv–2402, 2024.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=yBlVlS2Fd9.

Shengpeng Ji, Ziyue Jiang, Jialong Zuo, Minghui Fang, Yifu Chen, Tao Jin, and Zhou Zhao. Speech watermarking with discrete intermediate representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24239–24247, Apr. 2025b. doi: 10.1609/aaai.v39i23.34600. URL https://ojs.aaai.org/index.php/AAAI/article/view/34600.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 22605–22623, 2024.

Lauri Juvela and Xin Wang. Audio codec augmentation for robust collaborative watermarking of speech synthesis. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888976.

Slavko Kovačević, Murilo Z Silvestre, Kosta Pavlović, Petar Nedić, and Igor Djurović. Deepmark benchmark: Redefining audio watermarking robustness. In *The 1st Workshop on GenAI Watermarking*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Jason Li. Spectral codecs: Improving non-autoregressive speech synthesis with spectrogram-based audio codecs. *arXiv preprint arXiv:2406.05298*, 2024.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023.

Chang Liu, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Dear: A deep-learning-based audio re-recording resilient watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13201–13209, 2023a.

Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking. *arXiv preprint arXiv:2312.03410*, 2023b.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8):1448–1461, 2024.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Alexandre Mauroy, Y Susuki, and Igor Mezic. *Koopman operator in systems and control*, volume 484. Springer, 2020.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech 2021*, pp. 2127–2131, 2021.

Yigitcan Özer, Woosung Choi, Joan Serrà, Mayank Kumar Singh, Wei-Hsiang Liao, and Yuki Mitsufuji. A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs? *arXiv preprint arXiv:2505.19663*, 2025.

Kosta Pavlović, Slavko Kovačević, Igor Djurović, and Adam Wojciechowski. Robust speech watermarking by a jointly trained embedder and detector using a dnn. *Digital Signal Processing*, 122: 103381, 2022.

Kosta Pavlović, Lazar Stanarević, Petar Nedić, Slavko Kovačević, and Igor Djurović. Aware: Audio watermarking with adversarial resistance to edits. *arXiv preprint arXiv:2510.17512*, 2025.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chu Yuan Zhang, and Junzuo Zhou. Fewer-token neural speech codec with time-invariant codes. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12737–12741. IEEE, 2024.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics*, 2002.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, volume 235, 2024.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations*.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25719–25733, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6db0903efdfe9b1bbafb015c10990b78-Paper-Conference.pdf.

Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. SNAC: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL https://openreview.net/forum?id=PFBF5ctj4X.

Colm Sloan, Naomi Harte, Damien Kelly, Anil C Kokaram, and Andrew Hines. Objective assessment of perceptual audio quality using visqolaudio. *IEEE Transactions on Broadcasting*, 63(4): 693–705, 2017.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio and Speech Language Processing*, 19(7):2125–2136, 2011.

Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Yaoxun Xu, Jianwei Yu, Hangting Chen, Zhiyong Wu, Xixin Wu, Dong Yu, Rongzhi Gu, and Yi Luo. Wake: Watermarking audio with key enrichment. *arXiv preprint arXiv:2506.05891*, 2025.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm).*, 2019.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *Advances in Neural Information Processing Systems*, 37:56802–56827, 2024.

Dongchao Yang, Rongjie Huang, Yuanyuan Wang, Haohan Guo, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. Codec does matter: exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 25697–25705, 2025.

In-Kwon Yeo and Hyoung Joong Kim. Modified patchwork algorithm: A novel audio watermarking scheme. *IEEE Transactions on speech and audio processing*, 11(4):381–386, 2003.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*, 2025a.

Zheqi Yuan, Yucheng Huang, Guangzhi Sun, Zengrui Jin, and Chao Zhang. Multiplexing neural audio watermarks. *arXiv preprint arXiv:2511.02278*, 2025b.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AF9Q8Vip84.

Shengkui Zhao, Zexu Pan, and Bin Ma. Clearervoice-studio: Bridging advanced speech processing research and practical deployment. In *Proc. Interspeech 2025*, pp. 2980–2984, 2025.

A PROOFS

A.1 PROOF OF THE EQUIVALENT ROBUSTNESS CONDITION

Theorem 3.4 (Equivalent Robustness Condition). Let $\mathbf{m}' = 2\mathbf{m} - \mathbf{1}$ be the watermark message mapped from $\{0,1\}^N$ to $\{-1,1\}^N$. The detector is robust at $\mathbf{x}_w$ if and only if:

$$\begin{cases} m_i' \sum_{k=1}^{r} w_{ik}\psi_k(\mathbf{x}_w) & > 0 \\ m_i' \sum_{k=1}^{r} \lambda_k w_{ik}\psi_k(\mathbf{x}_w) & > 0 \end{cases} \tag{17}$$

Proof. According to Definition 3.1, a system is robust iff

$$\begin{cases} \mathbb{I}[f_i(\mathbf{x}_w) > 0] = m_i \\ \mathbb{I}[f_i(\mathcal{A}(\mathbf{x}_w)) > 0] = m_i \end{cases} \tag{18}$$

For the first equation:

$$\begin{aligned} \mathbb{I}[f_i(\mathbf{x}_w) > 0] = m_i &\iff \text{sgn}(f_i(\mathbf{x}_w)) = m_i' \\ &\iff m_i' f_i(\mathbf{x}_w) > 0 \\ &\iff m_i' \sum_{k=1}^{r} w_{ik}\psi_k(\mathbf{x}_w) > 0 \end{aligned} \tag{19}$$

For the second equation:

$$\begin{aligned} \mathbb{I}[f_i(\mathcal{A}(\mathbf{x}_w)) > 0] = m_i &\iff \text{sgn}(f_i(\mathcal{A}(\mathbf{x}_w))) = m_i' \\ &\iff m_i' f_i(\mathcal{A}(\mathbf{x}_w)) > 0 \\ &\iff m_i' \sum_{k=1}^{r} w_{ik}\psi_k(\mathcal{A}(\mathbf{x}_w)) > 0 \end{aligned} \tag{20}$$

Substituting the attack decomposition in Equation (4), we have:

$$\mathbb{I}[f_i(\mathcal{A}(\mathbf{x}_w)) > 0] = m_i \iff m_i' \sum_{k=1}^{r} w_{ik}\psi_k(\mathcal{A}(\mathbf{x}_w)) > 0 \iff m_i' \sum_{k=1}^{r} \lambda_k w_{ik}\psi_k(\mathbf{x}_w) > 0 \tag{21}$$

Combining Equations (19) and (21) completes the proof. □

A.2 PROOF OF THE WEAK ROBUSTNESS CONDITION

Theorem 3.5 (Weak Robustness Condition). The watermark system is robust at $\mathbf{x}_w$ if 1) The original detection is correct; 2) The margin of the robust component exceeds the magnitude of the fragile component. Formally,

$$\begin{cases} m_i' f_i(\mathbf{x}_w) > 0 \\ m_i' f_{i,\parallel}(\mathbf{x}_w) > |f_{i,\perp}(\mathbf{x}_w)| \end{cases} \tag{22}$$

Proof. According to Definition 3.1, a system is robust iff

$$\begin{cases} \mathbb{I}[f_i(\mathbf{x}_w) > 0] = m_i \\ \mathbb{I}[f_i(\mathcal{A}(\mathbf{x}_w)) > 0] = m_i \end{cases} \tag{23}$$

For the first equation:

$$\begin{aligned} m_i' f_i(\mathbf{x}_w) > 0 &\iff \text{sgn}(f_i(\mathbf{x}_w)) = m_i' \\ &\iff \mathbb{I}[f_i(\mathbf{x}_w) > 0] = m_i \end{aligned} \tag{24}$$

To derive the second equation, we decompose the scoring function as in Equation (8):

$$f_i(\mathcal{A}(\mathbf{x}_w)) = f_{i,\parallel}(\mathbf{x}_w) + f_{i,\perp}(\mathbf{x}_w) \tag{25}$$

By multiplying both sides by m'_i, we obtain the condition for correct detection under attack:

$$\begin{aligned} m'_i f_i(\mathcal{A}(\mathbf{x}_w)) &= m'_i(f_{i,\parallel}(\mathbf{x}_w) + f_{i,\perp}(\mathbf{x}_w)) \\ &= m'_i f_{i,\parallel}(\mathbf{x}_w) + m'_i f_{i,\perp}(\mathbf{x}_w) \end{aligned} \tag{26}$$

Since $m'_i f_{i,\perp}(\mathbf{x}_w) \geq -|f_{i,\perp}(\mathbf{x}_w)|$, we have

$$m'_i f_i(\mathcal{A}(\mathbf{x}_w)) \geq m'_i f_{i,\parallel}(\mathbf{x}_w) - |f_{i,\perp}(\mathbf{x}_w)| \tag{27}$$

Substituting in the second condition $m'_i f_{i,\parallel}(\mathbf{x}_w) - |f_{i,\perp}(\mathbf{x}_w)| > 0$, we have:

$$m'_i f_i(\mathcal{A}(\mathbf{x}_w)) > 0 \iff \mathbb{I}[f_i(\mathcal{A}(x_w)) > 0] = m_i \tag{28}$$

This satisfies the second requirement in Definition 3.1, completing the proof. □

A.3 Proof of Strong Robustness Condition

The watermark system is robust against attack $\mathcal{A}$ if: 1) The original detection is correct; 2) The detector is a linear combination of eigenfunctions with $\lambda_k = 1$. Formally,

$$\begin{cases} m'_i \sum_{k \in I_\parallel} w_{ik} \psi_k(\mathbf{x}_w) > 0 \quad \forall k \in I_\parallel \\ w_{ik} = 0 \quad \forall k \in I_\perp \end{cases} \tag{29}$$

Proof. A detector satisfying the second condition in Equation (29) can be formulated as:

$$f_i(\mathbf{x}) = \sum_{k \in I_\parallel} w_{ik} \psi_k(\mathbf{x}) \tag{30}$$

The first condition in Equation (29) ensures that:

$$m'_i f_i(\mathbf{x}_w) = \sum_{k \in I_\parallel} m'_i w_{ik} \psi_k(\mathbf{x}_w) > 0 \tag{31}$$

Meanwhile, since $\lambda_k = 1$ for all $k \in I_\parallel$, we have:

$$\begin{aligned} f_i(\mathcal{A}(\mathbf{x}_w)) &= \sum_{k \in I_\parallel} \lambda_k w_{ik} \psi_k(\mathbf{x}_w) \\ &= \sum_{k \in I_\parallel} w_{ik} \psi_k(\mathbf{x}_w) \\ &= f_i(\mathbf{x}_w) \end{aligned} \tag{32}$$

Therefore, we have

$$m'_i f_i(\mathcal{A}(\mathbf{x}_w)) = m'_i f_i(\mathbf{x}_w) > 0 \tag{33}$$

which means that both inequalities in Definition 3.1 are satisfied, leading to the conclusion that the system is robust against attack $\mathcal{A}$. □

B Linear Example

We illustrate the theory derived in Section 3.2 using a linear example. Let the generator be $G(\mathbf{x}, \mathbf{m}) = W_m[\mathbf{x}^\top \ \mathbf{m}^\top]^\top$, the detector be $f_i(\mathbf{x}) = {\mathbf{w}_d^i}^\top \mathbf{x}$, and the attack be a linear transformation $A(\mathbf{x}) = W_a \mathbf{x}$. This toy example demonstrates that eigenfunctions serve as a generalization of eigenvectors, and that robustness is fundamentally determined by the detector's alignment with the attack operator's principal subspace.

B.1 Connection with General Cases

In the linear example, each left eigenvector of W_a induces an eigenfunction of Koopman operator $\mathcal{K}_a$.

Lemma B.1. *Let $W_a \in \mathbb{R}^{T\times T}$ be a diagonalizable matrix. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$ be a set of linearly independent left eigenvectors of W_a, i.e., $\mathbf{u}_k^\top W_a = \lambda_k \mathbf{u}_k^\top$, $k = 1, \ldots, T$. Define the linear observable function $\psi_k(\mathbf{x}) = \mathbf{u}_k^\top \mathbf{x}$. Then:*

1. *Each ψ_k is an eigenfunction of the Koopman operator $\mathcal{K}_A$ with eigenvalue λ_k.*
2. *$\{\psi_1, \ldots, \psi_T\}$ forms a basis of linear function space $\mathcal{F}_{lin} = \{f : \mathbb{R}^T \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}\}$. In other words, every linear eigenfunction of $\mathcal{K}_A$ can be written uniquely as a linear combination of the ψ_k.*
3. *The mapping $\mathbf{u}_k \mapsto \psi_k$ is a bijection.*

B.2 Robustness Conditions

To derive linear version of robust conditions, we first decompose the linear detector vector $\mathbf{w}_d^i$ into linear combination of left eigenvectors $\{\mathbf{u}_k\}_{k=1}^T$ as in Equation (5). Specifically, let $\mathbf{w}_d^i = \sum_{k=1}^T w_{ik}\mathbf{u}_k$ where w_{ik} are the scalar coefficients. Then the detector score $f_i(\mathbf{x})$ can be written as:

$$f_i(\mathbf{x}_w) = {\mathbf{w}_d^i}^\top \mathbf{x}_w = \sum_{k=1}^T w_{ik}\mathbf{u}_k^\top \mathbf{x}_w = \sum_{k=1}^T w_{ik}\psi_k(\mathbf{x}_w) \tag{34}$$

where $\mathbf{x}_w$ represents the watermarked audio.

Similar to Equation (6), applying the attack $A(\mathbf{x}_w)$ modulates these components by eigenvalues λ_k:

$$f_i(A(\mathbf{x}_w)) = \sum_{k=1}^T \lambda_k w_{ik}\psi_k(\mathbf{x}_w) = \sum_{k=1}^T \lambda_k w_{ik}(\mathbf{u}_k^\top \mathbf{x}) \tag{35}$$

Then the linear version of the equivalent robustness condition in Theorem 3.4 is:

$$\begin{cases} m_i' \sum_{k=1}^T w_{ik}\mathbf{u}_k^\top \mathbf{x}_w & > 0 \\ m_i' \sum_{k=1}^T \lambda_k w_{ik}\mathbf{u}_k^\top \mathbf{x}_w & > 0 \end{cases} \tag{36}$$

Analogous to Equation (34), we split the term $f_i(A(\mathbf{x}_w))$ into a robust term $f_{i,\|}(\mathbf{x}_w)$ and a fragile term $f_{i,\perp}(\mathbf{x}_w)$:

$$f_i(A(\mathbf{x}_w)) = \underbrace{\sum_{k\in I_\|} w_{ik}\mathbf{u}_k^\top \mathbf{x}_w}_{f_{i,\|}(\mathbf{x}_w)} + \underbrace{\sum_{k\in I_\perp} \lambda_k w_{ik}\mathbf{u}_k^\top \mathbf{x}_w}_{f_{i,\perp}(\mathbf{x}_w)} \tag{37}$$

Taking Equation (37) into Equations (10) and (11), we can derive linear form of weak and strong robustness conditions.

B.3 Robustness Condition under the Optimal Attack

In this section, we simplify the robustness conditions derived in Section B.2 by considering the optimal linear compression attack that best preserves the original audio signal when rank of W_a is limited under r.

Lemma B.2 (Low-Rank Approximation)**.** *Let $\Sigma = \mathbb{E}_{\mathbf{x}_w}[\mathbf{x}_w\mathbf{x}_w^\top]$ be the covariance matrix of the watermarked signal. We decompose the covariance matrix into $\Sigma = \sum_{j=1}^T \sigma_j^2 \mathbf{u}_j\mathbf{u}_j^\top$ where the eigenvalues are sorted as $\sigma_1^2 \geq \cdots \geq \sigma_T^2$. Then W_a that minimizes reconstruction error $\mathbb{E}_{\mathbf{x}_w}[\|A(\mathbf{x}_w) - \mathbf{x}_w\|^2]$ s.t. rank$(A) \leq r \leq T$ is the orthogonal projection onto the principal subspace $\mathcal{S}_r = span\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$. i.e.:*

$$\arg \min_{\substack{W_a\in\mathbb{R}^{T\times T} \\ rank(W_a)\leq r}} \mathbb{E}_{\mathbf{x}_w}[\|A(\mathbf{x}_w) - \mathbf{x}_w\|^2] = P_{\mathcal{S}_r} = \sum_{j=1}^r \mathbf{u}_j\mathbf{u}_j^\top \tag{38}$$

which indicates that the eigenvalues for this attack are:

$$\lambda_k = \begin{cases} 1 & \text{if } 1 \le k \le r \quad (\textit{robust set } I_\parallel) \\ 0 & \text{if } r < k \le T \quad (\textit{fragile set } I_\perp) \end{cases}$$

Note that the fragile term vanishes when $W_a = P_{\mathcal{S}_r}$, since all eigenvalues in the fragile set equal to zero.

Theorem B.3 (Weak Robustness Condition)**.** *Decompose the detection vector $\mathbf{w}_d^i$ into a preserved component $\mathbf{w}_\parallel^i = P_{\mathcal{S}_r}\mathbf{w}_d^i$ and a discarded component $\mathbf{w}_\perp^i = (I - P_{\mathcal{S}_r})\mathbf{w}_d^i$. Then the linear system is robust against optimal attack $P_{\mathcal{S}_r}$ at $\mathbf{x}_w$ if and only if:*

$$m_i'({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) > \begin{cases} |{\mathbf{w}_\perp^i}^\top \mathbf{x}_w| & \text{if } sgn({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) \neq sgn({\mathbf{w}_\perp^i}^\top \mathbf{x}_w) \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

A simpler sufficient condition is that the magnitude of the robust score exceeds the magnitude of the fragile score:

$$m_i'({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) > |{\mathbf{w}_\perp^i}^\top \mathbf{x}_w| \tag{40}$$

Additionally, a stronger condition guarantees the above condition by constraining:

$${\mathbf{w}_\perp^i}^\top \mathbf{x}_w = 0 \tag{41}$$

Furthermore, the strong robustness condition guarantees robustness by forcing $\mathbf{w}_\perp^i = 0$:

Theorem B.4 (Strong Robustness Condition)**.** *The watermark system is robust to attack A if the system is valid on unattacked sample $\mathbf{x}_w$, and the detector vector resides in the principal subspace $\mathcal{S}_r$, i.e.:*

$$\begin{cases} m_i' \displaystyle\sum_{k=1}^{T} w_{ik}\mathbf{u}_k^\top \mathbf{x}_w > 0 \\ \qquad\qquad \mathbf{w}_d^i \in \mathcal{S}_r \end{cases} \tag{42}$$

Note that if we set $\mathcal{X} = \mathbb{R}^T$,

$$\mathbf{w}_d^i \in \mathcal{S}_r \iff f_i(\mathbf{x}) = f_i(A(\mathbf{x})), \quad \forall \mathbf{x} \in \mathcal{X} \tag{43}$$

However, assuming $\mathcal{X} = \mathbb{R}^T$ is not reasonable since there always exists an adversarial example $\mathbf{x}^*$ causing incorrect detection. Specifically, let $W_m = [W_x | W_{msg}]$, where $W_x \in \mathbb{R}^{T\times T}$ is the part of the matrix acting on the original audio. Whenever $W_x^\top \mathbf{w}_d^i \neq \mathbf{0}$, we can construct $\mathbf{x}^* = -\alpha m_i'(W_x^\top \mathbf{w}_d^i)$. As long as α is sufficiently large, the final watermarked signal $\mathbf{x}_w^* = W_x\mathbf{x}^* + W_{msg}\mathbf{m}$ will lead to an incorrect detection result. Formally,

$$f_i(\mathbf{x}_w^*) = {\mathbf{w}_d^i}^\top (W_x\mathbf{x}^* + W_{msg}\mathbf{m}) = -\alpha m_i' \|W_x^\top \mathbf{w}_d^i\|^2 + {\mathbf{w}_d^i}^\top W_{msg}\mathbf{m} \tag{44}$$

Since $m_i' \in \{-1, 1\}$, we have $(m_i')^2 = 1$. Therefore,

$$m_i' f_i(\mathbf{x}_w^*) = -\alpha \|W_x^\top \mathbf{w}_d^i\|^2 + m_i'({\mathbf{w}_d^i}^\top W_{msg}\mathbf{m}) \tag{45}$$

For a sufficiently large α, we have $m_i' f_i(\mathbf{x}_w^*) < 0$, which implies $\mathbb{I}[f_i(\mathbf{x}_w^*) > 0] \neq m_i$.

B.4 Proofs

Lemma B.1 Let $W_a \in \mathbb{R}^{T\times T}$ be a diagonalizable matrix. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$ be a set of linearly independent left eigenvectors of W_a, i.e., $\mathbf{u}_k^\top W_a = \lambda_k \mathbf{u}_k^\top$, $k = 1, \ldots, T$. Define the linear observable function $\psi_k(\mathbf{x}) = \mathbf{u}_k^\top \mathbf{x}$. Then:

1. Each ψ_k is an eigenfunction of the Koopman operator $\mathcal{K}_A$ with eigenvalue λ_k.
2. $\{\psi_1, \ldots, \psi_T\}$ forms a basis of linear function space $\mathcal{F}_{\text{lin}} = \{f : \mathbb{R}^T \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}\}$. In other words, every linear eigenfunction of $\mathcal{K}_A$ can be written uniquely as a linear combination of the ψ_k.

3. The mapping $\mathbf{u}_k \mapsto \psi_k$ is a bijection.

Proof. 1. For any $\mathbf{x} \in \mathbb{R}^T$,

$$(\mathcal{K}_A \psi_k)(\mathbf{x}) = \psi_k(A(\mathbf{x})) = \mathbf{u}_k^\top W_a \mathbf{x} = \lambda_k \mathbf{u}_k^\top \mathbf{x} = \lambda_k \psi_k(\mathbf{x})$$

2. First, for all k, $\psi_k \in \mathcal{F}_{\text{lin}}$. To show linear independence, assume $\sum_{k=1}^T c_k \psi_k = 0$ (the zero function). Then for every $\mathbf{x} \in \mathbb{R}^T$, $\sum_{k=1}^T c_k \mathbf{u}_k^\top \mathbf{x} = 0$. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$ be the set of right eigenvectors of W_a corresponding to $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$. Since W_a is diagonalizable, left and right eigenvectors satisfy the biorthogonality property $\mathbf{u}_k^\top \mathbf{v}_j = 0$ for $k \neq j$. Choosing $\mathbf{x} = \mathbf{v}_j$ (with $j = 1, \ldots, T$) yields $c_j = 0$. Hence the functions are linearly independent. Since the dimension of $\mathcal{F}_{\text{lin}}$ is T, the set $\{\psi_1, \ldots, \psi_T\}$ forms a basis.

3. Injectivity: if $\mathbf{u}_i \neq \mathbf{u}_j$, then choosing $\mathbf{x} = \mathbf{v}_i$ gives $\psi_i(\mathbf{v}_i) = \mathbf{u}_i^\top \mathbf{v}_i = 1$ while $\psi_j(\mathbf{v}_i) = \mathbf{u}_j^\top \mathbf{v}_i = 0$, namely $\psi_i \neq \psi_j$.
Surjectivity: $\{\psi_1, \ldots, \psi_T\}$ is precisely the image of the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$.
Hence the mapping is a bijection.

□

Theorem B.3 (Weak Robustness Condition). Decompose the detection vector $\mathbf{w}_d^i$ into a preserved component $\mathbf{w}_\parallel^i = P_{\mathcal{S}_r} \mathbf{w}_d^i$ and a discarded component $\mathbf{w}_\perp^i = (I - P_{\mathcal{S}_r}) \mathbf{w}_d^i$. Then the linear system is robust against optimal attack $P_{\mathcal{S}_r}$ at $\mathbf{x}_w$ if and only if:

$$m_i'({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) > \begin{cases} |{\mathbf{w}_\perp^i}^\top \mathbf{x}_w| & \text{if } \operatorname{sgn}({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) \neq \operatorname{sgn}({\mathbf{w}_\perp^i}^\top \mathbf{x}_w) \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

A simpler sufficient condition is that the magnitude of the robust score exceeds the magnitude of the fragile score:

$$m_i'({\mathbf{w}_\parallel^i}^\top \mathbf{x}_w) > |{\mathbf{w}_\perp^i}^\top \mathbf{x}_w| \tag{47}$$

Additionally, a stronger condition guarantees the above condition by constraining:

$${\mathbf{w}_\perp^i}^\top \mathbf{x}_w = 0 \tag{48}$$

Proof. In the linear system with optimal attack $P_{\mathcal{S}_r}$, we have:

$$f_i(\mathbf{x}_w) = \sum_{k=1}^r w_{ik} \mathbf{u}_k^\top \mathbf{x}_w + \sum_{k=r+1}^T w_{ik} \mathbf{u}_k^\top \mathbf{x}_w \tag{49}$$

$$= \sum_{k=1}^r {\mathbf{w}_d^i}^\top \mathbf{u}_k \mathbf{u}_k^\top \mathbf{x}_w + \sum_{k=r+1}^T {\mathbf{w}_d^i}^\top \mathbf{u}_k \mathbf{u}_k^\top \mathbf{x}_w \tag{50}$$

$$= (P_{\mathcal{S}_r} \mathbf{w}_d^i)^\top \mathbf{x}_w + [(I - P_{\mathcal{S}_r}) \mathbf{w}_d^i]^\top \mathbf{x}_w \tag{51}$$

$$= {\mathbf{w}_\parallel^i}^\top \mathbf{x}_w + {\mathbf{w}_\perp^i}^\top \mathbf{x}_w \tag{52}$$

Therefore,

$$m_i' f_i(\mathbf{x}_w) > 0 \iff m_i' {\mathbf{w}_\parallel^i}^\top \mathbf{x}_w > -m_i' {\mathbf{w}_\perp^i}^\top \mathbf{x}_w \tag{53}$$

For the second line of Equation (10), we have:

$$f_\parallel(\mathbf{x}_w) = \sum_{k=1}^r w_{ik} \mathbf{u}_k^\top \mathbf{x}_w = \sum_{k=1}^r {\mathbf{w}_d^i}^\top \mathbf{u}_k \mathbf{u}_k^\top \mathbf{x}_w = \sum_{k=1}^r (P_{\mathcal{S}_r} \mathbf{w}_d^i)^\top \mathbf{x}_w = {\mathbf{w}_\parallel^i}^\top \mathbf{x}_w \tag{54}$$

$$f_\perp(\mathbf{x}_w) = 0 \tag{55}$$

Therefore,

$$m_i' f_\parallel(\mathbf{x}_w) > |f_\perp(\mathbf{x}_w)| \iff m_i' {\mathbf{w}_\parallel^i}^\top \mathbf{x}_w > 0 \tag{56}$$

Combining Equations (53) and (56), we have:

$$m_i'(\mathbf{w}_{\parallel}^{i\top}\mathbf{x}_w) > \begin{cases} |\mathbf{w}_{\perp}^{i\top}\mathbf{x}_w| & \text{if } \operatorname{sgn}(\mathbf{w}_{\parallel}^{i\top}\mathbf{x}_w) \neq \operatorname{sgn}(\mathbf{w}_{\perp}^{i\top}\mathbf{x}_w) \\ 0 & \text{otherwise} \end{cases} \tag{57}$$

which is exactly Equation (39). □

C Implementation Details

C.1 Dataset Structure

The ID, architecture name, configuration name and partition of each codec are listed in Table 6.

C.2 Stable Function Approximation

The audio encoder is pretrained within the Vocos (Siuzdak, 2024) autoencoder framework to learn a robust latent representation. The configurations for audio processing, encoder pretraining and dimension selection are summarized in Table 4.

Table 4: Hyperparameters for Audio Processing, Encoder Pretraining, and USI Selection.

Audio Processing		**Pretraining**		**Dimension Selection**	
Sampling Rate	16000 Hz	Optimizer	AdamW	Invariance Weight α	0.2
Frame Length T	1024	Initial Learning Rate	5e-4	Transfer Weight β	0.2
FFT Size	1024	Batch Size	48	Cohesion Weight γ	0.4
Hop Size	256	Total Epochs	94	Consistency Weight δ	0.2
Number of Mel Filters	128	Loss Weight λ_D	1.0	Selected Dimension r	46
Latent Dimension d	128	Loss Weight λ_{mel}	45		

C.3 Watermark Embedding

The watermark embedding process involves iteratively updating the STFT amplitude spectrogram via backpropagation to minimize $\mathcal{L}_{wm}$. Specific optimization settings and the detection loss weight used during this phase are listed in Table 5.

Table 5: Watermark Embedding Hyperparameters.

Parameter	**Value**
Embedding Optimizer	NAdam
Learning Rate	0.1
Total Iterations	400
Message Loss Weight λ	0.1
Perturbation Threshold τ_{dB}	6 dB

C.4 VAD

To avoid watermark embedding and detection in silent segments, we integrate WebRTC VAD as a preprocessing module. This module partitions the input audio into 10 ms frames and utilizes a Gaussian Mixture Model (GMM) with an aggressiveness setting of 3 to classify each frame as voiced or unvoiced. Then we add a 20ms padding on both sides of each segment. After that, neighboring voiced segments with an interval of less than 30ms are merged. Finally, fragments with a duration of less than 50ms are discarded, and the remaining voiced segments are concatenated as the input of the audio encoder.

Table 6: Model configurations and dataset partitioning of neural codecs used in our experiments.

ID	Codec Name	Config	Partition
D1	EnCodec (Défossez et al., 2022)	6b24k	Train
D2	EnCodec (Défossez et al., 2022)	24b24k	Dev
B	DAC (Kumar et al., 2023)	DAC24k	
F1	FunCodec (Du et al., 2024)	en-libritts-16k-gr1nq32ds320	
I1	WavTokenizer (Ji et al., 2025a)	small_320_24k_4096	
J	Language-Codec (Ji et al., 2024)	language-codec	
W	SQ-Codec (Yang et al., 2025)	sqcodec50dim9	
A	SpeechTokenizer (Zhang et al., 2024)	hubert_avg	Test
C	HiFi-Codec (Yang et al., 2023)	16k_320d_large_universal	
E1	SemantiCodec (Liu et al., 2024)	1.40kbps_16k	
E2		0.70kbps_16k	
E3		0.35kbps_16k	
F2	FunCodec (Du et al., 2024)	zh_en-general-16k-nq32ds320	
G	FACodec (Ju et al., 2024)	encodec-decoder-v2_16k	
H	LLM-Codec (Yang et al., 2024)	llm_codec	
I2	WavTokenizer (Ji et al., 2025a)	medium_320_24k_4096	
K	HILCodec (Ahn et al., 2024)	hilcodec_speech	
L1	Vocos (Siuzdak, 2024)	vocos_encodec_6	
L2		vocos_encodec_12	
M	SoCodec (Guo et al., 2024)	16384x4_120ms_16khz_chinese	
P1	SNAC (Siuzdak et al., 2024)	24khz	
P2		44khz	
Q1	X-Codec (Ye et al., 2025)	hubert_librispeech	
R	AudioDec (Wu et al., 2023)	audiodec_24k_320d	
S	BigCodec (Xin et al., 2024)	BigCodec	
T1	TiCodec (Ren et al., 2024)	ticodec_1g1r	
T3		ticodec_1g4r	
U	Mimi (Défossez et al., 2024)	mimi-codec	
V	Spectral Codecs (Langman et al., 2024)	spectral codecs	

C.5 Determining Threshold τ for Calculating Existence Accuracy

As mentioned in Section 5.1, a threshold τ is required for calculating binary accuracy of each watermarking system: A sample is classified as "watermarked" (positive) if its bitwise accuracy exceeds τ. For each watermarking system, τ is optimized on a validation set containing 500 audio samples from LibriSpeech dev-clean, and subsequently applied on the unseen testset.

Specifically, we conduct binary search in range $[\tau_{\min}, \tau_{\max}]$ to find a τ that achieves the highest accuracy on attacked development set. $\tau_{\min}$ is determined as the minimal threshold needed to achieve False Positive Rate (FPR) < 0.1 on unattacked development set. Similarly, $\tau_{\max}$ is determined as the maximum threshold needed to achieve False Negative Rate (FNR) < 0.1. These boundaries ensure that the watermarking system maintains high existence accuracy on unattacked audio, which can be supported by the "Clean Acc." reported in Table 1.

D Additional Results

D.1 Limitation of ℓ_2-based Invariance Loss

In this section, we compare our cosine similarity-based invariance loss with ℓ_2 loss that strictly optimize towards precisely-robust eigenfunctions. Specifically, we compare our method (StableMark + Cosine Similarity) with three baselines:

1. **AWARE**: random-initialized audio encoder without pretraining.

2. **StableMark w/o $\mathcal{L}_{inv}$**: randomly-initialized audio encoder for dimension selection.
3. **StableMark + ℓ_2)**: Replacing the cosine similarity loss with ℓ_2 loss $\mathbb{E}_{\mathbf{x}}[\|\hat{\Psi}(\mathcal{A}(\mathbf{x})) - \hat{\Psi}(\mathbf{x})\|_2$

We test all three systems on 200 samples from LibriSpeech test clean. The results in Table 7 well support our speculation in Section 4.1. ℓ_2-based loss only allows exact equivalence during training, restricting the learning of eigenfunctions with $\lambda \neq 1$. This leads to worse watermarking performance in all Train, Dev and Test sets. It is noted that even in the clean set without codec resynthesis attack, ℓ_2-based loss performs significantly worse.

Table 7: Bitwise accuracy (%) for different encoder pretraining and alignment configurations. Clean denotes evaluation without an attack; Train, Dev, and Test denote the corresponding codec partitions.

Configuration	$\mathcal{L}_{inv}$	Clean	Train	Dev	Test
AWARE	N/A	99.00	67.18	66.99	60.29
StableMark	$\times$	**99.62**	74.67	70.39	62.68
	ℓ_2	89.53	72.05	67.10	64.51
	Cosine similarity	98.70	**76.70**	**71.23**	**65.04**

D.2 Per-Attack Accuracy under Neural Codec attacks

Table 8: Attack-wise bit/existence accuracy (%) comparison between baselines and StableMark.

Codec	AudioSeal	WavMark	Timbre	Timbre-SA	AWARE	StableMark (ours)
			Train			
D1	63.48 / 51.24	53.83 / 50.75	63.22 / 54.00	**90.26 / 99.79**	69.42 / 82.82	82.00 / 96.06
			Dev			
D2	88.79 / 85.68	53.77 / 50.25	68.31 / 59.52	**94.47 / 99.98**	79.17 / 94.08	89.87 / 97.97
B	88.76 / 82.94	89.63 / 95.92	90.39 / **100.00**	97.80 / 99.96	**99.56** / 98.28	99.51 / 98.41
F1	48.71 / 50.00	52.91 / 50.02	88.85 / **99.87**	86.80 / 99.64	78.74 / 94.22	**93.62** / 98.20
I1	46.61 / 50.00	53.11 / 50.25	54.45 / 50.17	51.23 / 51.03	54.93 / 55.42	**59.97 / 63.67**
J	40.67 / 50.00	**53.35** / 50.19	47.11 / 50.00	53.25 / 50.34	49.42 / 51.26	52.69 / **51.95**
W	47.43 / 49.94	45.53 / 49.75	45.77 / 50.00	**51.91 / 50.02**	50.32 / 49.20	48.52 / 49.18
			Test			
A	46.98 / 50.00	53.15 / 50.13	57.35 / 50.73	64.32 / 66.97	60.10 / 63.67	**69.07 / 82.23**
C	46.79 / 50.04	53.44 / 50.29	58.90 / 50.82	**67.87** / 74.57	58.36 / 60.65	65.62 / **75.30**
E1	46.35 / 50.00	55.34 / 50.71	66.61 / 55.88	**67.23 / 73.44**	53.96 / 53.52	59.39 / 63.08
E2	46.20 / 50.00	55.51 / 50.56	66.62 / 55.76	**67.25 / 73.61**	54.09 / 53.91	59.33 / 62.89
E3	46.55 / 50.00	55.69 / 50.50	66.55 / 55.92	**67.35 / 73.73**	53.78 / 53.85	59.53 / 62.14
F2	45.86 / 50.02	52.90 / 50.17	73.62 / 71.64	77.52 / 93.22	78.83 / 93.34	**87.73 / 97.82**
G	45.25 / 50.00	52.68 / 50.02	60.28 / 51.09	65.80 / 69.71	63.48 / 70.78	**72.13 / 87.88**
H	45.55 / 50.00	54.38 / 50.33	50.35 / 50.02	61.78 / 61.37	56.32 / 57.16	**63.11 / 70.49**
I2	47.47 / 54.96	53.16 / 50.13	57.24 / 51.72	54.42 / 52.53	55.53 / 56.01	**60.51 / 64.75**
K	64.73 / 50.00	53.67 / 50.06	76.47 / 78.72	**92.28 / 99.39**	74.98 / 90.58	86.99 / 97.61
L1	45.46 / 50.00	52.83 / 49.87	62.40 / 52.53	78.35 / 94.62	69.56 / 83.03	**80.38 / 95.25**
L2	45.38 / 50.00	52.80 / 50.13	64.83 / 53.29	80.64 / 96.96	74.54 / 90.79	**85.83 / 97.61**
M	48.18 / 50.00	56.75 / 50.79	53.67 / 50.29	**58.81** / 54.94	50.89 / 51.51	54.87 / **55.88**
P1	44.97 / 50.00	53.23 / 50.04	52.34 / 50.02	54.51 / 51.63	55.52 / 55.92	**61.15 / 66.07**
P2	51.31 / 51.09	53.10 / 50.15	68.10 / 59.90	**73.76 / 86.88**	61.39 / 65.91	70.28 / 84.62
Q1	46.17 / 49.94	53.00 / 49.75	52.01 / 50.00	54.23 / **50.02**	56.22 / 49.20	**61.83** / 49.18
R	48.55 / 49.98	54.43 / 50.27	57.32 / 52.24	63.43 / 65.63	61.58 / 65.93	**74.13 / 89.85**
S	46.96 / 50.00	53.06 / 50.34	51.67 / 50.13	57.98 / 55.94	60.07 / 63.33	**67.95 / 80.45**
T1	46.04 / 50.00	53.48 / 50.23	53.43 / 50.10	51.33 / 51.26	51.43 / 51.26	**55.81 / 56.64**
T3	47.57 / 50.02	53.92 / 50.08	50.42 / 50.27	63.33 / 64.25	58.97 / 61.37	**65.99 / 76.39**
U	47.60 / 50.02	54.59 / 50.08	61.79 / 53.49	**64.08** / 66.41	57.57 / 58.41	63.92 / **72.10**
V	47.36 / 50.00	52.91 / 50.02	74.98 / 77.29	80.21 / 96.09	85.58 / 97.41	**93.64 / 98.30**
Test Mean	47.60 / 50.28	53.82 / 50.21	60.77 / 55.54	66.66 / 71.51	61.49 / 65.80	**69.05 / 76.66**

As a complement to Figure 2, we report per-attack accuracy under neural codec attacks in numerical form in Tables 8 and 9. Results in Table 9 suggest that each architectural choice is indispensable, as removing any single component leads to a consistent performance decline across diverse neural codec attacks.

Table 9: Per-attack bit/existence accuracy (%) under neural codec attacks in ablation studies.

Dataset	w/o pretrain	w/o inv.	w/o recon.	w/o dim. sel.	w/o VAD	StableMark (ours)
			Train			
D1	56.46 / 50.00	74.32 / 87.19	51.75 / 52.36	76.48 / 94.41	77.72 / 92.72	**82.00 / 96.06**
			Dev			
D2	63.75 / 57.85	84.86 / 96.36	55.36 / 60.78	84.57 / 97.95	86.35 / **98.54**	**89.87** / 97.97
B	90.11 / 97.15	98.22 / 99.21	65.61 / 84.58	97.59 / 99.20	**99.64 / 99.77**	99.51 / 98.41
F1	66.62 / 62.37	89.47 / 98.07	54.45 / 57.81	88.60 / 98.77	92.38 / **99.64**	**93.62** / 98.20
I1	46.75 / 48.91	53.22 / 52.47	46.34 / 44.50	56.18 / 55.57	56.86 / 53.50	**59.97 / 63.67**
J	43.66 / 48.56	45.56 / 50.40	42.09 / 43.20	46.97 / 49.58	51.50 / 50.34	**52.69 / 51.95**
W	47.62 / 49.85	**49.38** / 49.90	48.66 / 33.61	44.68 / 47.57	49.35 / **50.25**	48.52 / 49.18
			Test			
A	49.78 / 48.79	61.59 / 62.45	47.85 / 45.08	64.21 / 73.57	66.08 / 70.01	**69.07 / 82.23**
C	48.88 / 48.72	58.32 / 57.43	47.90 / 45.35	60.73 / 64.63	62.32 / 61.74	**65.62 / 75.30**
E1	45.24 / 49.04	52.75 / 52.36	45.57 / 43.78	55.52 / 55.07	56.61 / 52.97	**59.39 / 63.08**
E2	45.24 / 48.95	52.68 / 52.39	45.59 / 43.91	55.62 / 55.04	56.71 / 53.04	**59.33 / 62.89**
E3	45.42 / 48.77	52.60 / 51.95	45.67 / 43.99	55.71 / 54.90	56.47 / 53.41	**59.53 / 62.14**
F2	64.08 / 58.67	82.58 / 95.38	52.24 / 54.37	81.89 / 97.03	84.66 / **97.84**	**87.73** / 97.82
G	52.32 / 49.08	65.92 / 71.12	48.52 / 47.55	67.26 / 80.39	68.35 / 75.97	**72.13 / 87.88**
H	47.43 / 48.93	55.57 / 54.50	46.40 / 45.14	58.88 / 60.84	59.48 / 56.68	**63.11 / 70.49**
I2	46.62 / 48.91	54.11 / 53.29	46.60 / 43.62	56.85 / 56.47	57.26 / 53.73	**60.51 / 64.75**
K	60.86 / 53.54	80.53 / 94.52	52.33 / 53.83	80.38 / 96.67	84.46 / **98.26**	**86.99** / 97.61
L1	56.81 / 50.17	73.81 / 86.25	50.74 / 50.00	74.86 / 93.22	76.55 / 92.17	**80.38 / 95.25**
L2	60.51 / 53.01	80.07 / 93.85	52.10 / 52.18	79.86 / 96.74	82.55 / 97.36	**85.83 / 97.61**
M	43.10 / 48.43	48.00 / 50.86	44.31 / 42.21	51.08 / 51.21	54.86 / 51.32	**54.87 / 55.88**
P1	47.27 / 48.99	54.31 / 53.73	46.54 / 44.12	57.53 / 57.95	59.16 / 56.19	**61.15 / 66.07**
P2	51.20 / 48.83	63.14 / 65.82	48.67 / 46.61	64.64 / 73.84	66.59 / 71.01	**70.28 / 84.62**
Q1	47.38 / 49.85	54.96 / 49.90	46.75 / 33.61	58.39 / 47.57	60.10 / **50.25**	**61.83** / 49.18
R	52.14 / 48.74	66.25 / 72.10	47.36 / 45.46	67.88 / 82.25	70.23 / 80.72	**74.13 / 89.85**
S	50.14 / 48.87	60.74 / 60.69	48.05 / 45.69	62.94 / 69.80	64.14 / 65.15	**67.95 / 80.45**
T1	44.12 / 48.89	48.69 / 50.86	44.92 / 42.68	51.43 / 50.88	53.52 / 51.36	**55.81 / 56.64**
T3	48.90 / 49.06	58.75 / 57.89	47.95 / 45.98	61.02 / 65.61	62.57 / 62.08	**65.99 / 76.39**
U	48.41 / 48.77	56.27 / 55.42	47.25 / 44.85	59.93 / 63.10	61.07 / 58.58	**63.92 / 72.10**
V	69.58 / 67.75	89.16 / 98.28	55.51 / 60.25	87.52 / 98.64	92.20 / **99.46**	**93.64** / 98.30
Test Mean	51.16 / 50.67	62.31 / 65.50	48.13 / 46.38	64.28 / 70.25	66.18 / 68.60	**69.05 / 76.66**

D.3 Per-Attack Accuracy under AI attacks

To demonstrate the generalization ability of the proposed method, we conduct evaluation on 4 types of AI attacks included in Deepmark Benchmark (Kovačević et al.). In addition to the average accuracy under AI attacks reported in Table 1, we further report the accuracy for each individual AI-based attack:

1. **Neural Vocoder** that resynthesizes the watermarked audio with BigVGAN (Lee et al., 2023).
2. **VAE** that reconstructs watermarked audio through a VAE named RAVE (Caillon & Esling, 2021).
3. **Diffusion Attack**: that regenerates the audio by first adding noise and then denoises with a DDPM-based audio diffusion model.
4. **Speech Enhancement**: that applies speech enhancement algorithm from ClearerVoice-Studio (Zhao et al., 2025).

Results in Table 10 show that our method achieves near-perfect bit accuracy under Neural Vocoder and Diffusion attacks, while surpassing most baselines under VAE and Speech Enhancement attacks.

D.4 Robustness against Audio Edits

Besides neural codec attacks and AI attacks, we additionally evaluate our system's robustness against conventional audio edits. Our evaluation is conducted on the strict version of RawBench (Özer et al., 2025), a standard benchmark testing the robustness of audio watermarking systems against a wide spectrum of audio edit attacks.

The results listed in Table 11 show that our system achieves robustness comparable to state-of-the-art baselines, maintaining a high mean bit accuracy of 95.22% across eighteen conventional audio edits. We observe relatively lower performance under dynamic operations, such as Limiter (LM),

Table 10: Per-attack bit/existence accuracy (%) under AI attacks. NV: Neural Vocoder; DFA: Diffusion Attack; SE: Speech Enhancement.

Codec	NV	VAE	DFA	SE	Mean
AudioSeal	57.23 / 52.16	41.35 / 50.00	49.54 / 50.06	87.03 / 83.70	58.79 / 58.98
WavMark	55.28 / 50.33	53.19 / 50.04	65.32 / 54.46	57.49 / 51.78	57.82 / 51.65
Timbre	**100.00 / 100.00**	52.29 / 50.17	**99.92 / 100.00**	79.77 / 85.87	83.00 / 84.01
Timbre-SA	96.10 / **100.00**	**62.28 / 62.56**	94.01 / 99.62	**98.77 / 100.00**	87.79 / **90.55**
AWARE	98.69 / 98.47	54.26 / 54.88	96.85 / 98.79	97.42 / 98.09	86.81 / 87.56
StableMark (ours)	99.47 / 98.16	57.84 / 60.07	99.57 / 98.49	95.69 / 98.76	**88.14** / 88.87
w/o pretrain	87.02 / 95.81	47.15 / 48.85	91.45 / 98.07	84.91 / 92.95	77.63 / 83.92
w/o inv. loss	98.09 / 99.23	50.21 / 51.53	98.30 / 99.33	91.99 / 98.58	84.65 / 87.17
w/o recon. loss	63.70 / 81.16	46.37 / 44.54	64.33 / 82.31	62.81 / 79.84	59.30 / 71.96
w/o dim. selection	97.45 / 99.46	54.11 / 53.56	97.43 / 99.46	90.83 / 98.95	84.96 / 87.86
w/o VAD	99.61 / 99.81	57.19 / 54.08	95.88 / 99.46	93.32 / 99.23	86.50 / 88.15

Table 11: Bit Accuracy under Audio Edits (%). (GN: Gaussian Noise; BN: Background Noise; RV: Reverberation; DC: Dynamic Range Compression; DE: Dynamic Range Expansion; LM: Limiter; LP: Lowpass Filter; HP: Highpass Filter; EQ: Equalizer; TS: Time Stretch; TJ: Time Jitter; PI: Polarity Inversion; GA: Gain; QN: Quantization; PS: Pitch Shift; MP: MP3 Compression; OG: Vorbis Compression; AA: AAC Compression)

Method	GN	BN	RV	DC	DE	LM	LP	HP	EQ	TS	TJ	PI	GA	QN	PS	MP	OG	AA	MEAN
AudioSeal	83	**100**	89	98	95	95	**100**	**100**	**100**	82	99	**100**	98	**100**	**100**	42	90	**100**	92.83
WavMark	99	**100**	87	99	99	99	**100**	96	91	97	**100**	19	99	**100**	62	**100**	**100**	**100**	91.50
Timbre	97	**100**	99	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99	**100**	**100**	**99.72**
Timbre-SA	99	**100**	95	99	99	99	99	99	99	99	99	**100**	99	**100**	**100**	98	**100**	**100**	99.06
AWARE	**100**	**100**	**100**	**100**	86	**100**	**100**	**100**	**100**	82	**100**	**100**	**100**	**100**	**100**	99	**100**	**100**	98.17
StableMark (ours)	99	99	98	88	85	77	99	97	99	81	99	99	98	99	**100**	98	99	**100**	95.22
w/o pretrain	99	99	98	85	75	73	99	97	99	79	99	99	97	99	99	98	99	99	94.00
w/o inv. loss	99	99	98	88	84	77	99	97	**99**	82	**99**	**100**	98	99	99	98	**99**	99	95.17
w/o recon. loss	77	78	71	61	54	57	77	71	76	56	77	77	73	77	74	74	**75**	77	71.22
w/o dim. sel.	98	99	97	87	81	75	99	96	98	81	99	99	97	99	99	98	99	99	94.44
w/o VAD	99	**100**	99	**100**	98	99	**100**	**100**	**100**	81	**100**	**100**	**100**	**100**	**100**	99	**100**	**100**	98.61

Dynamic Range Expansion (DE), and Dynamic Range Compression (DC). This robustness drop is probably induced by the use of VAD, which makes our system sensitive to energy modification. Nevertheless, it is worth noticing that this level of robustness is achieved without incorporating any simulated audio edits during pretraining, demonstrating the generalization ability of our method to conventional signal processing distortions.

D.5 Performance under Higher Capacity Settings

Table 12 evaluates the scalability of the proposed system as the message payload increases from 60 to 200 bits per sample. Although all models exhibit an expected decline in robustness due to increased information density, our method maintains a competitive edge in robustness against neural codec attacks. Meanwhile, our audio quality remains stable throughout the scaling process. These results suggest that the approximated eigenfunctions have sufficient volume to support high-density information embedding.

D.6 Generalization Ability of the Dimension Selection Algorithm

To verify the generalization ability of our learned eigenfunctions, we evaluate the distribution of distances between clean and attacked features using the Wasserstein-1 (W_1) distance as a proxy for stability. As illustrated in Figure 5, the selected Stable Subspace consistently yields lower distances than the discarded Residual Space across Train, Dev, and Test partitions, indicating that our selection method effectively captures fundamental acoustic invariants rather than overfitting to training codecs.

Table 12: System performance under higher payload capacities.

	Neural Codec Attacks			AI Attacks	Imperceptibility		
	Train	Dev	Test		PESQ↑	VISQOL↑	NISQA↑
60 bits							
Timbre	65.65	68.07	**63.10**	81.54	3.942	3.811	3.924
AWARE	61.34	63.04	56.16	79.83	3.946	3.845	3.604
StableMark (ours)	**70.24**	**68.57**	60.45	**83.11**	3.951	3.699	3.583
100 bits							
Timbre	64.04	65.73	58.41	80.2375	3.779	3.760	3.815
AWARE	56.29	58.37	52.41	72.11	4.128	3.984	3.726
StableMark (ours)	**67.7**	**67.24**	**60.07**	**80.45**	3.962	3.709	3.612
200 bits							
Timbre	58.33	62.19	55.91	**75.48**	3.681	3.552	3.714
AWARE	57.48	59.48	54.14	72.00	3.892	3.799	3.572
StableMark (ours)	**61.93**	**62.97**	**56.67**	74.57	3.946	3.655	3.643

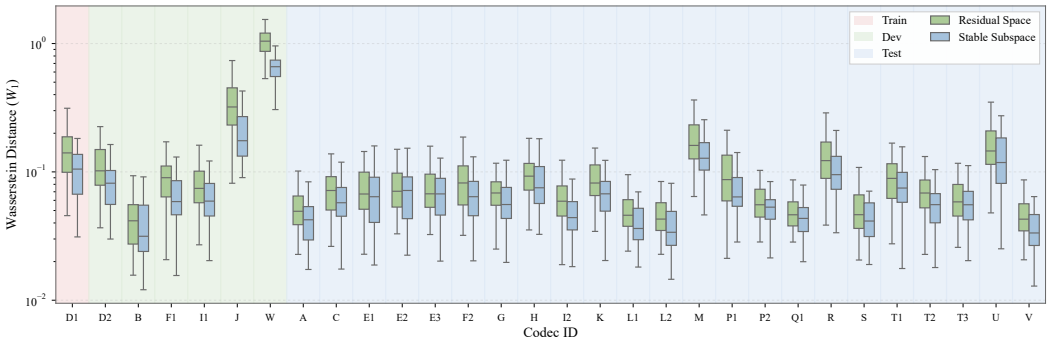

Figure 5: Distribution of Wasserstein-1 distances between clean and attacked features for the selected Stable Subspace and the Residual Space. The consistent lower distance in the Stable Subspace across Train, Dev, and Test partitions demonstrates that our dimension selection algorithm successfully identifies eigenfunctions that generalize to unseen codec architectures.

D.7 Computational Cost

In Table 13, we evaluate computational efficiency of our system by measuring Real-Time Factor (RTF) on a single NVIDIA A100 GPU. Here RTF is defined as the average time required to process a second of audio. All RTFs are tested with a batch size of 100 audios. Results indicate that StableMark supports real-time deployment. Meanwhile, the efficiency of our proposed method is comparable to feed-forward baselines like WavMark. Note that for detection stage, the real-time performance of embedder-free methods is significantly better than the feed-forward method. This demonstrates embedder-free method's advantage of utilizing only a light-weight watermark detector, which enables higher degree of parallelism for deployment.

Table 13: Computational cost for watermark embedding and detection in terms of (RTF).

Stage	WavMark	AWARE	StableMark (ours)
Embedding	8.60e-3	2.35e-2	2.65e-2
Detection	1.72e-1	5.26e-5	5.40e-5
VAD	-	-	3.990e-04

E LIMITATIONS

Despite its effectiveness, our approach has several limitations. First, our study focuses on speech watermarking under neural codec resynthesis, leaving broader audio domains and transformation families less explored. Second, StableMark is designed as a multi-bit post hoc audio watermarking method. Its extension to model watermarking, where watermarks are embedded during audio generation, remains an open direction. Finally, while StableMark prioritizes robustness, trade-off among robustness, watermark capacity, and perceptual quality may also need to be further investigated.