LATENT-MARK: An Audio Watermark Robust to Neural Codec Compression

Yen-Shan Chen[1,2,*], *Shih-Yu Lai*[1,3,4,*], *Ying-Jung Tsou*[1], *Yi-Cheng Lin*[1], *Bing-Yu Chen*[1], *Yun-Nung Chen*[1], *Hung-yi Lee*[1,5], *Shang-Tse Chen*[1,**]

[1]National Taiwan University, Taiwan [2]CyCraft AI Lab, Taiwan
[3]RIKEN Center for Computational Science (RIKEN-CCS), Japan
[4]MoonShine Animation Studio, Taiwan
[5]NTU Artificial Intelligence Center of Research Excellence (NTU AI-CoRE)

{r14922018, r13922a22, r14922076}@csie.ntu.edu.tw, f12942075@ntu.edu.tw, robin@ntu.edu.tw, y.v.chen@ieee.org, hungyilee@ntu.edu.tw, stchen@csie.ntu.edu.tw

Abstract

While existing audio watermarking techniques have achieved strong robustness against traditional digital signal processing (DSP) attacks, they remain vulnerable to neural compression. This occurs because modern neural audio codecs act as noise filters and discard the imperceptible waveform variations used in prior watermarking methods. To address this limitation, we propose Latent-Mark, the first zero-bit audio watermarking framework designed to survive neural codec compression. Our key insight is that robustness to the encode-decode process requires embedding the watermark within the codec's invariant latent space. We achieve this by optimizing the audio waveform to induce a detectable directional shift in its encoded latent representation, while constraining perturbations to align with the natural audio manifold to ensure imperceptibility. To prevent overfitting to a single codec's quantization rules, we introduce Cross-Codec Optimization, jointly optimizing the waveform across multiple surrogate codecs to target shared latent invariants. Extensive evaluations demonstrate robust zero-shot transferability to unseen neural codecs, achieving competitive resilience against traditional DSP attacks while preserving perceptual imperceptibility. We hope our work will inspire future research into universal watermarking frameworks capable of maintaining integrity across increasingly complex and diverse generative distortions. [1]

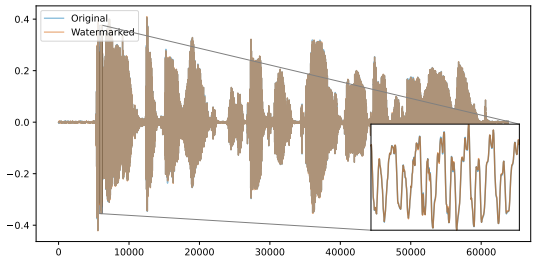
Index Terms: Audio Watermarking, Neural Codec Compression, Latent-Space Shift, Manifold Alignment, Cross-Codec Transferability

1. Introduction

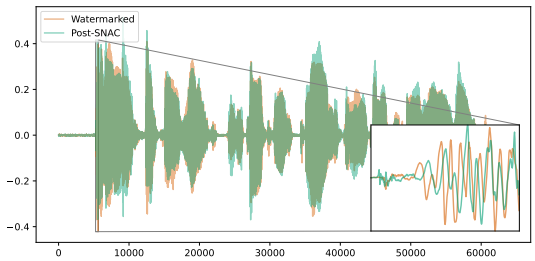
Audio watermarking has emerged as a critical tool for intellectual property protection. Recent state-of-the-art methods, such as AudioSeal [1], WavMark [2], and Timbre [3], demonstrate strong resilience against a broad spectrum of conventional digital signal processing (DSP) distortions [4], including compression, filtering, resampling, and sample suppression. In such settings, watermark robustness has largely been framed as surviving waveform- or spectrogram-level perturbations while remaining imperceptible to human listeners.

However, the rapid adoption of neural network-based codec compression introduces a qualitatively different threat surface that breaks this traditional notion of robustness. Modern neural audio codecs, such as EnCodec [5] and SNAC [6], reconstruct audio by mapping waveforms into discrete latent tokens and decoding them back under a strict bitrate constraint. This encode–quantize–decode pass, which we refer to as **Neural Codec Compression**, is not a localized DSP distortion but a highly non-linear projection through a learned latent bottleneck. As a result, existing watermarks that are highly robust to DSP transformations can fail catastrophically after a single codec pass. We investigated this impact through a preliminary analysis using AudioSeal [1], the SNAC [6] codec, and the LibriSpeech [7] dataset. As illustrated in Figure 1a, while a watermarked waveform initially exhibits no visible deviation from the original signal, a single neural encode–decode pass (Figure 1b) introduces extensive phase shifts and amplitude distortions that completely misalign with the source watermark.

We attribute this behavior to the inherent design of neural codecs, which act as a projector mapping input signals onto the latent space of valid audio representations. Under this projection,



(a) *Original vs. Watermarked*



(b) *Watermarked vs. Post-SNAC*

Figure 1: *Waveform comparisons at different processing stages using AudioSeal. (a) The original and watermarked waveforms largely overlap, showing minimal differences. (b) Substantial amplitude distortion and phase shifts occur after the SNAC encoding–decoding process.*

*These authors contributed equally.
**indicates the corresponding author.
[1]All source code and implementation details are available at: https://github.com/yenshan0530/Latent-Mark

traditional watermarks—typically embedded as imperceptible noise—are treated as off-manifold residuals and discarded during reconstruction. Despite the output remaining perceptually transparent, these structural modifications effectively strip away the fine-grained signals of watermarks required for detection. This elevates neural codec compression from a routine compression task to a potent watermark removal attack, representing a pressing real-world vulnerability as codec-based pipelines become the de facto standard for generative modeling and audio distribution.

To withstand this regime, we propose Latent-Mark, the first *zero-bit* audio watermarking framework (i.e., it encodes only *presence* rather than a payload) explicitly designed to withstand neural codec compression. Our core insight is that only features embedded in a codec's invariant latent space can survive its encode–quantize–decode process. Accordingly, we formulate watermark embedding as a latent-targeted optimization problem: we apply gradient-based updates directly to the input waveform to induce a detectable *directional shift* in the codec latent space, while constraining waveform perturbations so that the watermark remains imperceptible. By embedding the mark into the latent manifold itself, it becomes a feature the codec is designed to preserve, rather than a waveform-level artifact neural codecs are trained to discard.

A practical watermark must satisfy more than surviving neural codec distortions. **First**, it must remain imperceptible. We ensure this by constraining the latent shift to align with the codec's learned representation space—specifically along directions defined by codebook centroids. This alignment leverages the decoder to naturally regularize the watermark, preserving acoustic fidelity. **Second**, it must generalize beyond a single codec. Building on the above white-box embedding strategy, we further introduce cross-codec optimization across diverse surrogate codecs. By enforcing the watermark objective simultaneously under multiple quantization rules, the framework captures shared latent patterns across different codecs and avoids overfitting to a single architecture. This ensures zero-shot transferability to unseen black-box codecs. **Finally**, the watermark should remain robust to traditional DSP distortions: in addition to neural codec compression, we evaluate Latent-Mark and find that it maintains state-of-the-art resilience against a wide range of analytic attacks, including adding Gaussian noise, scaling amplitude, filtering, and resampling.

Our primary contributions are as follows:

- We identify **neural codec compression** as a fundamentally different attack regime for audio watermarking, and argue that neural codecs act as manifold projectors that can erase the imperceptible noise patterns used by prior watermarking methods.
- We propose Latent-Mark, the first zero-bit audio watermarking framework explicitly designed to withstand neural codec compression by inducing a detectable **latent directional shift** via gradient-based waveform optimization. We show that **latent space-aligned** shifts enhance acoustic **imperceptibility** (Sections 5.1, 5.3).
- Building on the white-box formulation, we introduce cross-codec optimization across multiple codecs to achieve strong zero-shot **transferability** to unseen black-box codecs. Ultimately, Latent-Mark provides a balance of perceptual fidelity and survivability to neural codec compression, while maintaining highly competitive **robustness against prior DSP attacks** (Section 5.4).

2. Related Work

2.1. Mechanisms of Neural Watermarking

Across audio, image, text, and multi-modal domains, watermarking methods are categorized by the space in which they embed signals [8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. A recurring theme is that robustness depends on how watermark signals interact with model-internal representations and survive distortions. To formalize this, we categorize several core concepts commonly employed in the design of robust watermarking algorithms:

Post-Hoc Audio Signal and Additive Perturbations. While traditional watermarking treats signals as additive perturbations across modalities [12], audio presents a uniquely stringent setting due to the ubiquity of lossy compression and model-mediated transformations [18]. In contrast to image or text modalities, audio watermarks face the unique challenge of surviving the complex digital signal processing (DSP) inherent to audio production workflows, followed by subsequent degradation across distribution channels. Specifically, they face repeated neural codec re-synthesis and quantization (e.g., SNAC [6], APCodec [19], and FunCodec [20], etc.), which RAW-Bench [9] identifies as the most formidable challenge to bit-string integrity. Furthermore, Özer *et al.* [9] emphasize that practical audio watermarks must withstand an extensive battery of studio-standard manipulations—including *dynamic range compression, limiting, equalization, and reverberation*—alongside environmental degradation like background noise. Audio watermarking methods like AudioSeal [1], SilentCipher [21], Timbre [3], and WavMark [2] address these vulnerabilities by incorporating psychoacoustic masking and noise-to-mask ratio losses [22] to hide information within inaudible frequency bands.

Latent and Manifold-Aware Embedding. Recognizing that deep generative models manipulate semantic concepts rather than raw signals, recent work has shifted toward embedding marks directly into continuous latent spaces or discrete codec tokens [11, 23, 10]. Building on classical manifold learning [24, 25] and vector-quantized representations [26], these methods inject structured perturbations into deep features to find a subspace separable from natural content variations. In this paradigm, a watermark's survival depends on aligning with model-internal structural invariances—such as pitch or speaker identity—rather than superficial signal details [27, 28]. To ensure robustness against the quantization and re-synthesis inherent in model-mediated transformations, current strategies include training-time codec augmentation [29] or integrating watermarking objectives directly into neural codecs via end-to-end joint training [10]. This motivates evaluating watermark schemes explicitly under latent representations, such as those used in multi-modal or diffusion-based frameworks [16, 17].

Audio Latent-Space Approaches and Their Limitations. Most recently, highly concurrent works have attempted to tackle the neural codec bottleneck by operating near or within latent spaces, though they rely on fundamentally different paradigms. For instance, Roman *et al.* [30] explore watermarking the *training data* of audio generative models such that the trained model inherently emits watermarked codec tokens. However, this assumes white-box control over the model training pipeline and fails to address the *post-hoc* marking of arbitrary audio assets in the wild. Alternatively, Liu *et al.* [31] propose an end-to-end framework utilizing cross-attention mechanisms for robust embedding. While effective against known distortions, it relies on training a static, feed-forward neural encoder against a prede-

fined attack set. This risks overfitting to the specific quantization rules seen during training, fundamentally limiting its *zero-shot transferability* to unseen, proprietary neural codecs. Furthermore, without explicit geometric constraints, modifying latent representations can easily push the signal off the natural audio manifold, introducing perceptible artifacts upon decoding. In contrast, our Latent-Mark overcomes these limitations. Rather than training a static encoder or altering generative models, we employ test-time *Cross-Codec Optimization* across surrogate codecs to induce a measurable directional shift in shared latent structures. This ensures zero-shot survival against black-box neural codec compression while enforcing natural manifold constraints to guarantee imperceptibility.

2.2. Watermark Robustness and the Threat of Distortions

The definition of a "robust" watermark has evolved significantly, shifting from algorithmic signal degradations to complete audio reconstruction.

Traditional DSP, Channel Distortions, and Robustness. Robustness is increasingly defined by rigorous evaluation protocols that transition from algorithmic signal-level corruptions to complex audio reconstructions. Historically, benchmarks like AudioMarkBench [8] and RAW-Bench [9] systematized a wide array of removal and forgery attacks, including MP3 compression, bandpass filtering, added noise, reverberation, and modern neural codec chains. To counter these, researchers have developed various defensive mechanisms, such as noise-to-mask ratio (NMR) losses [22], training-time codec augmentation [29], and end-to-end deepfake verification pipelines [32]. Beyond generic perturbations, contemporary systems must withstand adaptive and model-aware threats, including overwriting, ownership conflicts, and spoofing [33]. These security challenges have prompted the integration of defensive cryptographic hashing and filter-based mechanisms to prevent forgery within neural watermarking frameworks [34].

The Paradigm Shift to Neural Codec Compression. The landscape of distortions fundamentally changed with the introduction of high-fidelity neural audio codecs (e.g., SoundStream [35], EnCodec [5], DAC [36]). Unlike MP3, which removes psychoacoustically masked frequencies, neural codecs completely deconstruct the waveform into discrete tokens—typically utilizing convolutional architectures and Residual Vector Quantization (RVQ) to capture coarse-to-fine structure—and re-synthesize it from scratch. Because neural codec compression acts as an extreme, structure-aware information bottleneck [4, 37, 38], it explicitly discards the "off-manifold" residual noise that traditional additive methods rely upon. Compounding this threat, these discrete codec tokens now serve as the foundational vocabulary for modern generative sequence models [39, 40, 41]. Large audio language models treat quantized indices as linguistic sequences for zero- or few-shot speech synthesis [18, 42, 43, 44, 45]. Thus, signal-space modifications must survive this encode–quantize–decode pipeline. Recent benchmarks like RAW-Bench [9] confirm neural codecs are the primary threat to audio watermarks, necessitating a shift toward latent-aware strategies.

3. Methodology

3.1. Preliminaries

The Gap. The vulnerability of traditional audio watermarking to neural codec compression stems from a fundamental **representation mismatch**. Conventional methods typically embed watermarks as psychoacoustic masking patterns at the waveform level. However, modern neural codecs process audio through a lossy reconstruction operator:

$$\mathcal{R}(s) = \mathcal{D}(\mathcal{Q}(\mathcal{E}(s))) \tag{1}$$

where $\mathcal{E}$ is the encoder mapping the time-domain audio s to a continuous feature space; $\mathcal{Q}$ is the vector quantizer that discretizes these features by mapping them to the nearest entries in a codebook $\mathcal{W}$; and $\mathcal{D}$ is the decoder that reconstructs the signal back into the original audio space. Because traditional watermarks are designed to be imperceptible, they often manifest as subtle signal variations that are effectively treated as quantization noise. During codec compression, these "off-manifold" details are discarded. Consequently, the core challenge stems from **detectability**: the basic ability to extract the watermark from an unattacked signal, and **survivability**: the capacity of the watermark to endure the lossy quantization of codec compression.

The intuition. Assuming a white-box setting where we have parameter access to the codec's encoder $\mathcal{E}$, we propose Latent-Mark: a framework (Figure 2) that identifies and leverages latent properties invariant to the quantization process. Rather than injecting waveform-level noise, we induce a directional shift in the continuous latent representation z toward a secret manifold axis v_c before it enters the quantizer $\mathcal{Q}$. This modification is designed to survive codec compression by "steering" the latent tokens into a distribution with a persistent directional bias. While the perceived audio remains unchanged, this shift remains detectable even after the signal is decoded and subsequently re-encoded. We formulate this as a **zero-bit watermarking challenge**, where the goal is to induce a statistically significant shift in the latent sequence along a designated secret axis, ensuring the perturbation survives the neural compression bottleneck.

3.2. Problem Formulation

Let $s \in \mathbb{R}^T$ be the input audio waveform, where T denotes the total number of temporal samples. A neural codec c maps s to a latent representation:

$$z_c = \mathcal{Q}(\mathcal{E}(s)) \in \mathbb{R}^{d \times L} \tag{2}$$

where d is the codebook dimension and L is the sequence length of the latent frames. Our framework consists of an Encoder $\mathcal{E}_{wm}$ and a Detector $\mathcal{E}_{Det}$. We generate a watermarked waveform $s_{wm} = s + \delta$ by solving a gradient-based optimization over the perturbation δ. The Detector determines the presence of the mark by computing a scalar score y based on the latent projection of a suspect signal s' onto a secret manifold vector $v_c \in \mathbb{R}^d$.

To induce the desired shift while maintaining perceptual transparency, we formulate the embedding as a constrained optimization problem. We solve for the perturbation δ that maximizes the alignment with the manifold vector v_c subject to a small waveform distortion:

$$\min_{\delta} \mathcal{L}_{\text{wm}}(s + \delta, v_c) \quad \text{s.t.} \quad ||\delta||_\infty \leq \epsilon \tag{3}$$

where ϵ is a dynamic threshold determined by the target Signal-to-Distortion Ratio (SDR). This prevents the watermark from introducing audible artifacts into the original audio s.

The Latent-Mark framework consists of an optimization-based Encoder and a statistical Detector. For any given codec c, the presence of a watermark is defined by the alignment of the signal's latent representation z_c with a secret manifold axis

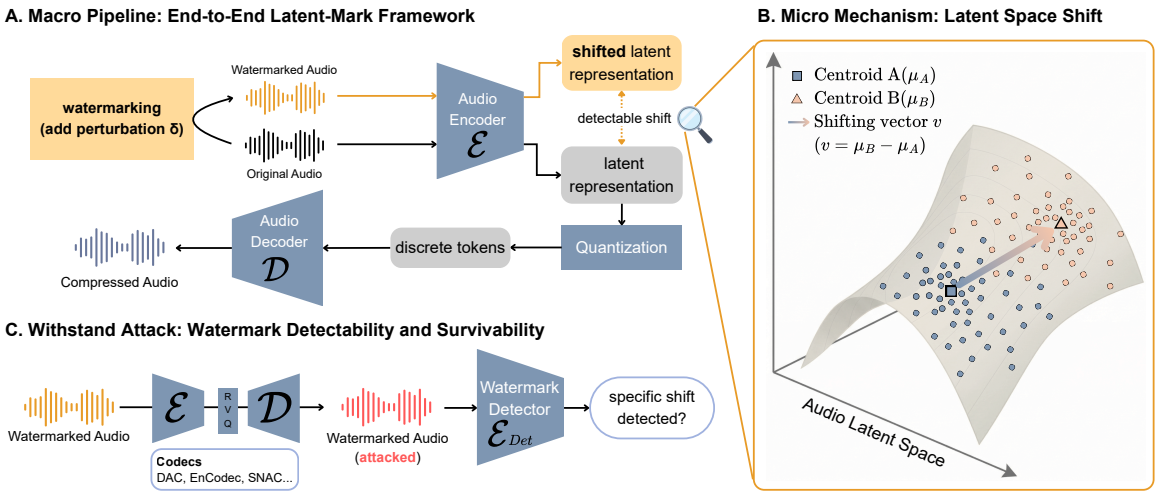


Figure 2: ***Overview of Latent-Mark.*** *(A)* ***Macro*** *end-to-end pipeline. The lower blue and gray components illustrate the standard neural codec compression pipeline; The orange components highlight our proposed additions: adding an optimized perturbation δ to the original audio to intentionally induce a constrained shift in its latent representation prior to quantization and decoding. (B)* ***Micro*** *mechanism for the robust latent space shift, driven by a predefined shifting vector $v = \mu_B - \mu_A$ between cluster centroids. (C) Watermark detectability and survivability against* ***attacks****. The watermarked audio undergoes destructive neural codec pipelines (encoding, RVQ, and decoding), after which a detector ($\mathcal{E}_{Det}$) verifies if the latent shift persists.*

$v_c \in \mathbb{R}^d$. We measure this alignment via the projection score $\bar{p}_c(s)$, defined as the temporal mean of the latent projections over the sequence length L:

$$\bar{p}_c(s) = \frac{1}{L}\sum_{t=1}^{L}\langle z_{c,t}, v_c\rangle \tag{4}$$

Watermark Optimization. Embedding is formulated as a constrained optimization problem. Given an input audio s, we solve for a waveform perturbation δ that steers the latent representation into the target direction. We minimize a hinge loss objective:

$$\min_{\delta} \mathcal{L}_{\text{wm}}(s+\delta, v_c) = \min_{\delta} \text{ReLU}(\gamma_c - \bar{p}_c(s+\delta)) \tag{5}$$

where γ_c represents the **target alignment score** (set to 1.5 in our implementation). The intuition behind γ_c is to create a "safety margin" that pushes the latent projection far enough into the target manifold so that the shift survives the rounding effects of the quantization bottleneck. To maintain imperceptibility, we enforce a dynamic threshold $\epsilon = \beta \cdot \text{RMS}(s) \cdot 10^{-\text{SDR}/20}$, where RMS denotes the root mean square function. We solve this via the Adam optimizer for 150 steps, applying a hard clip to δ at each iteration for imperceptibility.

Watermark Detection. Detection is a statistical verification process performed in the latent domain. For a suspect signal s', we first compute the raw projection score $\bar{p}_c(s')$, which corresponds to the temporal mean of the projected latent sequence. To ensure robustness across different codec architectures with varying latent scales, we compute a **Normalized Margin** m_c:

$$m_c(s') = \frac{\bar{p}_c(s') - \tau_c}{\sigma_c} \tag{6}$$

where τ_c is the detection threshold ($\mu_c + k\sigma_c$) and σ_c is the standard deviation of projections derived from a null distribution of clean audio. A watermark is detected if $m_c > 0$.

Choice of Shifting Axis. While v_c can be any arbitrary unit vector, for the main experiment, we design its selection with the intention of improving survivability through the quantization bottleneck $\mathcal{Q}$, and term this the **Latent-Cluster** variant. Aiming to guide the shift toward high-density regions of the latent space, we derive v_c by partitioning the codebook weights $W \in \mathbb{R}^{K\times d}$ into two primary groupings via K-means clustering ($k = 2$). Letting μ_A and μ_B be the resulting centroids, we define the axis as the unit-normalized vector between them:

$$v_c = \frac{\mu_B - \mu_A}{||\mu_B - \mu_A||_2} \tag{7}$$

By aligning v_c with the codebook distribution, our intention is for the perturbation to act more like a structural feature rather than stochastic noise, hypothesizing that this alignment increases its likelihood of preservation during discretization. A comparative analysis of alternative directions is in Section 5.1.

3.3. Cross-Codec Optimization

A fundamental limitation of single-codec optimization is the lack of transferability; a watermark optimized for one specific codec's latent representation may be treated as stochastic noise and discarded by another. To achieve zero-shot robustness against unseen black-box models, we introduce **Cross-Codec Optimization**. Instead of targeting a single bottleneck, we identify a directional shift that a committee of heterogeneous surrogate codecs $\mathcal{C}$ collectively deems structural.

Framework Overview. Our joint optimization pipeline consists of four integrated stages: (1) a multi-rate resampling loop to synchronize heterogeneous codec views; (2) a calibration phase to balance gradients across different latent scales; (3) a constrained optimization loop that induces a structurally invariant latent shift; and (4) an ensemble detection mechanism that aggregates evidence for robust verification.

Stage 1. Cross-Codec Resampling Pipeline. Because the committee members operate at different sampling rates, we implement a synchronized resampling loop. During each optimization step, the perturbation δ is maintained in a high-resolution

workspace at rate f_{work}. For each codec $c \in \mathcal{C}$, the perturbed signal $s + \delta$ is resampled to the codec's native rate f_c before latent extraction. Waveforms are padded to satisfy a temporal constraint T_{pad} to ensure stable optimization across all views. In our implementation, we set $f_{\text{work}} = 44.1\,\text{kHz}$ and T_{pad} to multiples of 4096 samples.

Stage 2. Gradient Balancing via Calibration. Different architectures operate on vastly different latent scales, which can lead to "gradient dominance" where one codec's loss overwhelms others. To counter this, we use a **Baseline Calibration** method to calculate a target threshold $\tau_c = \mu_c + k\sigma_c$ and a normalization scale α_c derived from a null distribution of clean audio:

$$\alpha_c = \mathbb{E}\left[\text{ReLU}(\tau_c - \bar{p}_c(s))\right] \tag{8}$$

where α_c represents the average projection gap of clean audio. We set the calibration constant $k = 1.5$. By dividing each per-codec hinge loss by its respective α_c, we ensure the optimization assigns equal importance to satisfying the manifolds of all committee members regardless of their latent variance.

Stage 3. Cross-Codec Optimization. We solve for the optimal perturbation δ by minimizing the ensemble normalized hinge loss over N_{steps} using the Adam optimizer:

$$\min_{\delta} \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\text{ReLU}(\tau_c - \bar{p}_c(s+\delta))}{\alpha_c} \quad \text{s.t.} \quad ||\delta||_\infty \leq \epsilon \tag{9}$$

The budget ϵ is dynamically adjusted as $\epsilon = \text{clip}(\beta \cdot \text{RMS}(s) \cdot 10^{-\text{SDR}/20}, \epsilon_{min}, \epsilon_{max})$. We set $N_{steps} = 150$, $\beta = 2.5$, $\epsilon_{min} = 10^{-4}$, and $\epsilon_{max} = 0.1$. This objective effectively induces a directional bias into the shared latent features of the committee, increasing the likelihood that the mark will survive codec compression by an unseen attacker codec $a \notin \mathcal{C}$.

Stage 4. Ensemble Detection and Δ-Score. Detection is performed by aggregating evidence across the committee. For a suspect signal s', we compute the **Normalized Margin** $m_c(s') = (\bar{p}_c(s') - \tau_c)/\sigma_c$ for each view. Let $\{m_{(1)}, m_{(2)}, \ldots, m_{(|\mathcal{C}|)}\}$ denote the margins sorted in ascending order. To ensure robustness against outlier distortions, the final detection score is defined as:

$$\text{score}(s') = m_{\left(\left\lceil \frac{|\mathcal{C}|}{2} \right\rceil\right)}(s') \tag{10}$$

Notably, taking the mean of normalized margins is susceptible to extreme outliers caused by scale variations across different codec geometries, while the median acts as a robust statistic. It effectively functions as a majority voting mechanism, ensuring that even if an attack completely obliterates the watermark in a minority of views (resulting in unbounded negative margins), the global detection remains stable as long as the structural bias persists in the remaining surrogate spaces.

To evaluate transferability under attack, we utilize a **Δ-Score** metric, which measures the shift in detection score relative to a clean baseline: $\Delta = \text{score}(R_a(s_{wm})) - \text{score}(R_a(s))$, where R_a is the codec compression operator of an unseen attack codec. A positive Δ indicates that the watermark's directional bias has successfully transferred through the black-box bottleneck.

4. Experiments

The primary objective of our study is to verify whether Latent-Mark retains its embedded signal through the extreme bottleneck of neural codec compression, and investigate if cross-codec joint optimization enables better transferability to unseen codecs.

4.1. Experimental Setup

Datasets. We evaluate our method across nine diverse datasets spanning three primary acoustic domains: *ambient and environmental sound* (AIR [46], Clotho [47]), *human speech* (LibriSpeech [7], DAPS [48]), and *music and vocals* (PCD [49], jaCappella [50], MAESTRO [51], GuitarSet [52], Freischuetz [53]). To ensure statistical consistency and balanced evaluation, we uniformly sample 120 instances from each dataset, randomly subsampling those that exceed this threshold.

Baselines. We compare Latent-Mark against three state-of-the-art watermarking models: **WavMark** [2], **SilentCipher** [21], and **AudioSeal** [1]. For the single-codec variant of Latent-Mark, we utilize SNAC as the primary surrogate model. For the cross-codec variant (**Latent-Mark-Joint**), we optimize across an ensemble of models with diverse architectures and sampling frequencies (SNAC 32 kHz, DAC 16 kHz, and DAC 44 kHz), testing cross-family codecs transferability on APCodec [19] and FunCodec [20], as formulated in Section 3.3. An extended analysis of this surrogate model selection is provided in Section 5.2.

Evaluation Metrics. We evaluate watermark performance using two primary metrics. **Detectability (Det.)** evaluates the detector's fundamental accuracy on clean, unperturbed audio. We assess watermark performance using a balanced test set comprising 60 watermarked and 60 unwatermarked (original audio) samples per dataset. Alongside overall classification accuracy, we explicitly report the True Positive Rate (TPR) and the False Positive Rate (FPR). **Survivability (Sur.)** quantifies the robustness of the embedded signal against severe compression bottlenecks. It is defined strictly as the successful detection rate on the watermarked samples after they have been processed through the neural codec compression attack.

Attacker Models. In these primary evaluations, we utilize the SNAC [6] architecture (with frequency 24kHz) as the surrogate model for Latent-Mark's detector. While our main experiments focus on survivability against SN codec compression, we further investigate the zero-shot transferability of our embedded marks in Section 5.2, where we test attacks against other unseen neural codecs in different codec types and sampling rates. We use default values for all hyperparameters.

4.2. Results

Table 1 (Left) summarizes the performance of our proposed Latent-Mark variants against state-of-the-art watermarking baselines across eight diverse datasets. From these evaluations, we draw three primary findings:

First, in terms of **detectability** on unperturbed audio, both the prior baselines and our Latent-Mark variants (with or without joint optimization) maintain highly competitive performance, consistently exceeding 0.95 accuracy across most datasets. This confirms that Latent-Mark, similar to prior baselines, reliably triggers the detector under standard conditions.

Second, a stark contrast emerges regarding **survivability** against the neural codec bottleneck. While prior watermarking methods experience catastrophic failure—dropping to near-zero detection rates across the board—Latent-Mark robustly preserves the embedded signal, achieving survivability scores consistently above 0.58 and peaking at 0.93 on the DAPS dataset (using Latent-Cluster). We note that *Latent-Joint* exhibits slightly lower performance compared to its single-codec counterparts; this is expected, as performing white-box optimization against multiple codecs simultaneously necessitates a trade-off in specialized

Table 1: *Benchmark results across datasets. Detectability (Det.) is presented as Accuracy (TPR/FPR), while Survivability (Sur.) reports the detection rate after neural codec compression with SNAC (24kHz). For simplicity, the original version of our **Latent-Mark** is shown as **Latent-Cluster**, and **Latent-Mark-Joint** is abbreviated to **Latent-Joint**. **Left:** Comparison of Latent-Mark and its joint-optimization variant against prior baselines (Section 4.2). **Right:** Evaluation of our method using alternative secret key directions (Section 5.1). Best and second-best values are highlighted in bold and underlined.*

		Main Experiment					*Directional Variants*	
Dataset	Metric	WavMark	SilentCipher	AudioSeal	Latent-Cluster	Latent-Joint	Latent-PCA	Latent-Random
AIR	Det.	95.0 (99.2/9.2)	93.3 (95.0/9.2)	<u>96.7</u> (97.5/4.2)	95.8 (95.8/4.2)	**100.0** (100.0/0.0)	95.8 (99.2/6.7)	**100.0** (100.0/0.0)
	Sur.	0.0	0.0	0.0	<u>61.7</u>	53.3	60.8	**66.7**
Clotho	Det.	93.3 (95.0/9.2)	**96.7** (100.0/5.8)	<u>94.2</u> (94.2/5.8)	92.5 (91.7/7.5)	**96.7** (95.8/3.3)	93.3 (92.5/7.5)	93.3 (95.0/9.2)
	Sur.	0.0	0.0	0.0	<u>58.3</u>	<u>58.3</u>	**61.7**	**61.7**
DAPS	Det.	**100.0** (100.0/0.0)	93.3 (97.5/11.7)	**100.0** (100.0/0.0)	<u>99.2</u> (100.0/1.7)	83.3 (85.0/20.0)	95.0 (99.2/9.2)	95.8 (96.7/5.0)
	Sur.	0.0	0.0	8.3	**93.3**	76.7	<u>88.3</u>	65.0
LibriSpeech	Det.	<u>96.7</u> (96.7/3.3)	<u>96.7</u> (97.5/4.2)	91.7 (91.7/7.5)	**100.0** (100.0/0.0)	93.3 (95.0/10.0)	95.8 (95.0/3.3)	95.0 (95.0/5.0)
	Sur.	4.2	0.0	5.0	**80.8**	74.2	65.8	<u>76.7</u>
jaCappella	Det.	<u>96.7</u> (96.7/3.3)	**100.0** (100.0/0.0)	**100.0** (100.0/0.0)	**100.0** (100.0/0.0)	95.0 (98.3/8.3)	95.8 (95.8/4.2)	**100.0** (100.0/0.0)
	Sur.	0.0	0.0	0.0	<u>75.8</u>	<u>75.8</u>	**77.5**	70.8
PCD	Det.	90.8 (90.8/9.2)	<u>96.7</u> (94.2/0.0)	90.8 (91.7/10.0)	90.8 (90.8/9.2)	88.3 (91.7/16.7)	95.8 (94.2/1.7)	**100.0** (100.0/0.0)
	Sur.	0.0	0.0	1.7	**81.7**	<u>78.3</u>	71.7	75.8
MAESTRO	Det.	<u>96.7</u> (99.2/5.0)	<u>96.7</u> (95.0/0.8)	95.0 (99.2/9.2)	**100.0** (100.0/0.0)	89.2 (95.0/16.7)	95.0 (98.3/8.3)	95.8 (95.0/3.3)
	Sur.	0.0	0.0	0.0	<u>80.8</u>	71.7	65.0	**83.3**
GuitarSet	Det.	**100.0** (100.0/0.0)	**100.0** (100.0/0.0)	**100.0** (100.0/0.0)	96.7 (95.0/0.8)	85.0 (90.0/20.0)	95.0 (95.0/5.0)	<u>97.5</u> (100.0/4.2)
	Sur.	0.0	0.0	0.0	**86.7**	68.3	<u>85.0</u>	61.7

robustness. Nevertheless, its performance remains highly competitive, maintaining accuracy rates above 0.58 across all tested configurations.

Finally, we observe that the audio domain influences retention. Speech and vocal datasets (e.g., DAPS, LibriSpeech) generally exhibit more stable survivability, frequently exceeding 0.70. In contrast, certain environmental noise distributions like AIR show slightly lower retention ranges (between 0.50 and 0.70). Notably, among the baselines, only AudioSeal demonstrated any marginal resilience to neural codec compression (retaining a mere 0.08 on DAPS and 0.05 on LibriSpeech), further underscoring the critical necessity of our latent-space formulation.

5. Ablation and Analysis

Having demonstrated the primary robustness and detectability of Latent-Mark under neural codec compression, we now present a series of analyses to validate our design. Our evaluation is twofold. First, we conduct **architectural ablations** to justify our core hyperparameter choices, specifically investigating the *selection of secret key directions* within the latent space (Section 5.1) and the *impact of surrogate model combinations* (Section 5.2) Second, we evaluate Latent-Mark against established **baselines** to compare the *acoustic imperceptibility* (Section 5.3) of the watermarks and verify that optimizing for neural codec bottlenecks does not inadvertently compromise *robustness against traditional digital signal processing attacks* (Section 5.4).

5.1. Choice of Secret Key Direction

We first discuss how the direction of the secret key (see Section 3.2, **Choice of Shifting Axis**) influences watermarking performance and survivability. To isolate the impact of this geometric choice, we compare three distinct variants for deriving the target axis: **Latent-Cluster** (our primary method, which utilizes the vector connecting the centroids of a $k = 2$ clustering of the codebook), **Latent-PCA**, which derives the axis from the first principal component (the direction of maximum variance) of the centered codebook weights using Singular Value Decomposition, and **Latent-Random**, which simply samples a uniformly random, unit-normalized vector within the latent dimension space.

Comparing the post-codec compression survivability of these approaches in Table 1 (Right), Latent-Cluster yields the most robust performance, ranking first in four of the eight datasets. Latent-Random follows closely, securing first place in three datasets and second place in one. Conversely, Latent-PCA consistently performs the worst across the evaluations.

We hypothesize that this hierarchy stems from how each strategy interacts with the quantization bottleneck $\mathcal{Q}$. For example, **Latent-Cluster** may explicitly guide the watermark towards denser regions of the codebook; by mimicking a structural transition between mass centers, the perturbation is effectively preserved by the nearest-neighbor quantizer as a valid latent shift. Conversely, the poor performance of **Latent-PCA** may reveal that shifting along the axis of maximum variance is detrimental. Because the first principal component represents the most continuous variations in the feature space, the quantizer likely treats such shifts as standard signal variance rather than a distinct structural feature, causing the watermark to be aggressively washed out during discretization.

Further comparisons regarding the acoustic imperceptibility of the watermarks generated by these distinct directional methods are detailed in Section 5.3.

5.2. Surrogate Model Selection for Joint Optimization

We next discuss how the model combination used for joint optimization (Section 3.3) affects performance and transferability.

Throughout this section, codec shorthand (e.g., `SNAC32`) denotes the architecture (`SNAC`) operating at a specific sampling frequency (`32` kHz). We select five optimization combinations

Table 2: *Transferability under codec compression attacks across all optimization settings and 3 acoustic domain datasets. Values are survivability pass rates (%). The best and second-best values per column are highlighted in* ***bold*** *and underlined.*

Method	*Ambient & Environmental*			*Human Speech*						*Music & Vocals*					
	Clotho			**LibriSpeech**			**DAPS**			**PCD**			**jaCappella**		
	SNAC44	*EnCodec48*	*DAC24*	*SNAC44*	*EnCodec48*	*DAC24*	*SNAC44*	*EnCodec48*	*DAC24*	*SNAC44*	*EnCodec48*	*DAC24*	*SNAC44*	*EnCodec48*	*DAC24*
Latent-Joint (Opt_C1)	92.50	97.50	93.33	**100.00**	85.00	**100.00**	90.83	**100.00**	**100.00**	95.00	**100.00**	90.00	92.00	**100.00**	68.00
Latent-Joint (Opt_C2)	**100.00**	**98.33**	79.17	80.00	**100.00**	79.17	**100.00**	**100.00**	77.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	98.00
Latent-Joint (Opt_F1)	79.17	95.83	57.50	**100.00**	91.67	**100.00**	98.33	**100.00**	69.17	77.50	**100.00**	97.50	96.00	**100.00**	88.00
Latent-PCA (SNAC24)	69.00	63.17	83.67	66.83	28.00	69.17	84.00	89.83	99.33	**100.00**	**100.00**	**100.00**	71.20	58.00	76.00
Latent-Cluster (SNAC24)	94.67	95.33	**97.16**	**100.00**	**100.00**	97.50	**100.00**	**100.00**	**100.00**	97.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Latent-Random (SNAC24)	66.00	55.33	75.33	76.83	54.17	83.33	88.50	53.67	99.00	2.50	0.00	10.00	77.20	44.00	98.00
Latent-Joint (Opt_D1)	60.00	55.80	59.20	65.67	70.80	73.33	79.20	95.80	72.50	37.50	37.50	50.00	60.00	62.00	80.00
Latent-Joint (Opt_D2)	70.80	63.30	61.70	47.50	69.17	44.17	90.80	**100.00**	82.50	55.00	52.50	52.50	40.00	44.00	56.00

targeting two orthogonal generalization axes: *cross-codec* architecture shifts and *cross-sampling-rate* variations.

- **Opt_C1 {SNAC32, DAC16, DAC44}** and **Opt_C2 {SNAC32, EnCodec24, EnCodec32}** focus on **intra-family** generalization. By optimizing against "architecturally closer" relatives sharing a hierarchical RVQ structure, we test if the watermark generalizes to unseen members of the same lineage across varying sampling rates.
- **Opt_F1 {SNAC24, DAC24, EnCodec24}** evaluates **cross-family** transferability under a controlled frequency constraint by fixing all codecs at `24` kHz.
- **Opt_D1 {SNAC24, DAC44, FunCodec}** and **Opt_D2 {SNAC24, FunCodec, APCodec}** explore extreme **cross-family** scenarios. These "distant" groups incorporate heterogeneous architectures (e.g., FunCodec, APCodec) to stress-test robustness against radical domain shifts in unrelated neural synthesis pipelines.

We compare these against single-optimization baselines (Latent-Cluster, PCA, and Random, optimized on **SNAC24**). To test transferability, we use `SNAC44`, `EnCodec48`, and `DAC24` for neural codec compression.

Experimental results are summarized in Table 2. We define **transferability** as the success rate of watermark perturbations—initially optimized to survive a specific codec compression—when evaluated against unseen codec architectures. Specifically, we first identify watermark perturbations that successfully survive `SNAC24` codec compression (the common optimization target across all settings) and subsequently assess whether this robustness generalizes to disparate codec environments. We present results for two representative datasets from each audio category: environmental noise, human speech, and music. Several key observations can be drawn from these results:

Architectural proximity serves as the primary determinant of transfer success, as evidenced by the significant performance gap between intra-family and cross-family transfers. Specifically, configurations optimized for architectures similar to the target (e.g., C_1, C_2, F_1) outperform distant-family transfers by approximately 20%. While identifying shared latent features across disparate architectures remains a challenge, our method maintains a baseline transferability between 50–70%, indicating that the learned perturbations are not strictly overfitted to a single decoder's bias. Furthermore, intra-family generalization is significantly increased by including at least one representative codec from the target family during optimization. For instance, the inclusion of DAC variants in C_1 leads to high robustness for the unseen `DAC24`, while EnCodec-inclusive optimization in C_2 excels on `EnCodec48`, suggesting that structural commonalities, such as shared Residual Vector Quantization (RVQ) hierarchies, are more critical for transfer than matching bitrates or sampling frequencies.

The joint optimization configurations consistently demonstrate superior robustness compared to single-variant baselines, provided that the optimization set includes at least one representative from the target codec's architectural family. By comparing the joint frameworks (C_1, C_2, F_1) against the single-variant baselines (*Latent-Cluster*, *PCA*, and *Random*), we observe that exposure to diverse neural codec compression processes prevents the watermark from occupying narrow, codec-specific latent regions, effectively covering a broader spectrum of unseen codecs than the individual strategies.

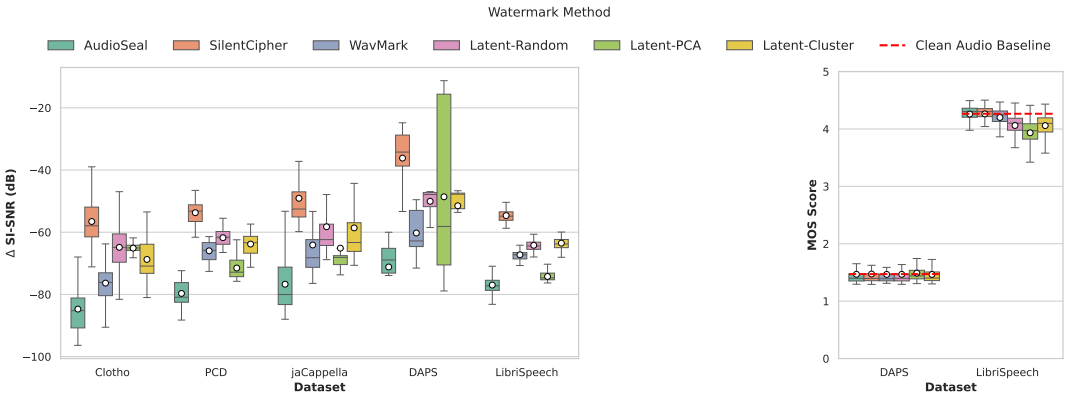
Finally, structural architecture remains a more significant bottleneck for watermark survival than sampling frequency, a fact highlighted by the F_1 configuration. Despite F_1 being optimized for 24kHz targets, it fails to provide disproportionate gains for the 24kHz `DAC24` variant compared to other families. This indicates that the specific inductive bias of the neural synthesis layers—rather than the frequency response of the signal—is the dominant factor that the watermark must navigate to remain imperceptible yet detectable.

5.3. Audio Quality and Imperceptibility

To evaluate the impact of watermark embedding on perceptual quality, we utilize two metrics: **ΔSI-SNR** (Scale-Invariant Signal-to-Noise Ratio change) [54] to measure mathematical waveform fidelity, and **UTMOS** (UTokyo-SaruLab MOS Prediction System) [55], a neural-based Mean Opinion Score (MOS) predictor for human-like perceptual assessment.

As illustrated in Figure 3a, the observed **change in waveform fidelity (ΔSI-SNR)** follows the trend: *SilentCipher* > *Latent-Cluster, Latent-Random* > *Latent-PCA*, *AudioSeal*, *WavMark*. Regarding our proposed variants, *Cluster* and *Random* demonstrate comparable stability, likely due to their uniform treatment of the latent space. In contrast, the PCA variant exhibits higher variance, reflecting greater sensitivity to the underlying data.

As for **perceptual quality (UTMOS)**, despite the differences captured by ΔSI-SNR, the UTMOS results in Figure 3b indicate that the perceptual quality across all methods is nearly indistinguishable, suggesting that the artifacts introduced by Latent-Mark are effectively masked and maintain a high level of imperceptibility to the human ear.



(a) *ΔSI-SNR across datasets.*

(b) *UTMOS on DAPS and LibriSpeech.*

Figure 3: *Comparison of objective waveform fidelity (ΔSI-SNR) and perceptual quality (UTMOS) across watermarking methods.*

Table 3: *Survivability (pass rates, %) under DSP attacks across datasets. Attacks: GAU = Gaussian noise (SNR=60), AMP = Amplitude scaling (0.5), LPF = Low-pass filtering (4kHz), and RSM = Resampling (16kHz).*

Dataset	Attack	AudioSeal [1]	SilentCipher [21]	WavMark [2]	**Latent-Mark**
AIR	GAU	**100.00**	18.60	3.33	**100.00**
	AMP	**100.00**	0.00	3.49	**100.00**
	LPF	**100.00**	10.17	3.49	**100.00**
	RSM	**100.00**	9.30	3.49	**100.00**
Freischuetz	GAU	68.49	97.26	**100.00**	69.86
	AMP	**100.00**	0.00	**100.00**	71.23
	LPF	**100.00**	93.15	**100.00**	56.16
	RSM	**100.00**	95.89	**100.00**	64.38
GuitarSet	GAU	90.00	**100.00**	**100.00**	67.78
	AMP	**100.00**	0.00	**100.00**	91.11
	LPF	**100.00**	94.44	**100.00**	**100.00**
	RSM	**100.00**	92.22	**100.00**	**100.00**
jaCappella	GAU	20.00	46.00	84.00	**100.00**
	AMP	**100.00**	0.00	84.00	88.00
	LPF	**100.00**	38.00	84.00	78.00
	RSM	**100.00**	34.00	84.00	84.00
LibriSpeech	GAU	**100.00**	99.19	**100.00**	**100.00**
	AMP	**100.00**	98.39	**100.00**	**100.00**
	LPF	**100.00**	89.52	**100.00**	75.81
	RSM	**100.00**	91.13	**100.00**	**100.00**

5.4. Robustness to Prior Attacks

Besides imperceptibility, we extend our comparisons to evaluate the robustness of Latent-Mark against traditional digital signal processing (DSP) distortions, benchmarking against the same three baselines: AudioSeal, WavMark, and SilentCipher. Following the evaluation protocol outlined in SoK [4], we subject the watermarked audio to four highly diverse signal distortion attacks to cover a wide spectrum of degradation: Gaussian noise, amplitude scaling, low-pass filtering, and resampling. The hyperparameters are set consistently with prior work.

As shown in Table 3, AudioSeal consistently achieves the highest robustness across these traditional distortions, which is expected given that it is explicitly trained on augmented audio editing data to ensure resilience against such modifications. Meanwhile, Latent-Mark and WavMark exhibit competitive, dataset-dependent performance; our method outperforms WavMark on the AIR and jaCappella datasets, whereas WavMark holds an advantage on Freischuetz and GuitarSet. This indicates that while Latent-Mark is primarily designed to survive neural codec bottlenecks, it successfully retains robustness against conventional attacks, performing on par with dedicated robust watermarking methods. Finally, SilentCipher yields the lowest detection rates in this setting—dropping to as low as 0.00% under amplitude scaling attacks across several datasets—which we attribute to its reliance on delicate temporal and phase alignments that are easily disrupted by broad signal distortions.

Summary of Trade-offs. Concluding from Sections 5.3 and 5.4, Latent-Mark establishes a balance between acoustic transparency and robustness. While AudioSeal and WavMark exhibit strong resilience to traditional DSP distortions, they do so at the severe expense of audio quality and completely fail under neural codec compression. Conversely, while SilentCipher maintains high imperceptibility, it remains vulnerable to both DSP attacks and codec bottlenecks. Latent-Mark delivers acoustic fidelity comparable to the highly constrained SilentCipher, while uniquely surviving the extreme bottleneck of neural codec compression and maintaining competitive defense against standard DSP perturbations.

6. Conclusion

We propose Latent-Mark, the first zero-bit audio watermarking framework specifically engineered to survive neural codec compression, a process that causes catastrophic failure in traditional methods. Using the insight that watermark information must be embedded directly into the latent space manifolds to survive quantization bottlenecks, we successfully bridge the gap between DSP robustness and neural codec survivability.

Our findings indicate that latent-space alignment is critical in robustness towards neural codec compression; specifically, perturbations guided by the codebook's topological clusters achieve higher imperceptibility while maintaining high detection accuracy. Beyond its specialized resilience to neural synthesis, *Latent-Mark* achieves competitive performance in both perceptual transparency and robustness against DSP attacks. We show that black-box transferability can be achieved by jointly optimizing across diverse codec architectures, and we hope our work will inspire future research on unified watermarking methods to adapt to the evolving landscape of generative neural synthesis.

7. Acknowledgment

This work was supported in part by the National Science and Technology Council under Grants NSTC 114-2634-F-002-004 and NSTC 114-2634-F-002-003-MBK.

8. References

[1] R. San Roman, P. Fernandez, H. Elsahar, A. D´efossez, T. Furon, and T. Tran, "Proactive detection of voice cloning with localized watermarking," *ICML*, 2024.

[2] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "WavMark: Watermarking for audio generation," 2023. [Online]. Available: https://arxiv.org/abs/2308.12770

[3] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via timbre watermarking," in *Network and Distributed System Security Symposium*, 2024.

[4] Y. Wen, A. Innuganti, A. B. Ramos, H. Guo, and Q. Yan, "SoK: How robust is audio watermarking in generative AI models?" 2025. [Online]. Available: https://arxiv.org/abs/2503.19176

[5] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," 2022. [Online]. Available: https://arxiv.org/abs/2210.13438

[6] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, "SNAC: Multi-scale neural audio codec," 2024. [Online]. Available: https://arxiv.org/abs/2410.14411

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[8] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Gong, "AudioMarkBench: Benchmarking robustness of audio watermarking," in *NeurIPS*, 2024.

[9] Y. Özer, W. Choi, J. Serrà, M. Singh, W.-H. Liao, and Y. Mitsufuji, "A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?" in *Proc. Interspeech 2025*, 2025, pp. 5113–5117.

[10] J. Zhou, J. Yi, Y. Ren, J. Tao, T. Wang, and C. Y. Zhang, "WMCodec: End-to-end neural speech codec with deep watermarking for authenticity verification," 2024. [Online]. Available: https://arxiv.org/abs/2409.12121

[11] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun, and H. Guan, "Attack-resilient image watermarking using stable diffusion," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[12] Y. Hu, Z. Jiang, M. Guo, and N. Z. Gong, "A transfer attack to image watermarks," in *The Thirteenth International Conference on Learning Representations*, 2025.

[13] A. Diaa, T. Aremu, and N. Lukas, "Optimizing adaptive attacks against watermarks for language models," 2025. [Online]. Available: https://arxiv.org/abs/2410.02440

[14] R. Chen, Y. Wu, J. Guo, and H. Huang, "De-mark: Watermark removal in large language models," in *Forty-second International Conference on Machine Learning*, 2025.

[15] S. Rastogi, P. Maini, and D. Pruthi, "STAMP your content: Proving dataset membership via watermarked rephrasings," in *Forty-second International Conference on Machine Learning*, 2025.

[16] J. Qiu, W. Han, X. Zhao, S. Long, C. Faloutsos, and L. Li, "Evaluating durability: Benchmark insights into multimodal watermarking," 2024. [Online]. Available: https://arxiv.org/abs/2406.03728

[17] S. Liu, Q. Zheng, J. J. Xu, Y. Yan, J. Zhang, H. Geng, A. Liu, P. Jiang, J. Liu, Y.-C. Tam, and X. Hu, "VLA-Mark: A cross modal watermark for large vision-language alignment model," 2025. [Online]. Available: https://arxiv.org/abs/2507.14067

[18] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[19] Y. Ai, X.-H. Jiang, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "AP-Codec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 3256–3269, Jun. 2024.

[20] Z. Du, S. Zhang, K. Hu, and S. Zheng, "FunCodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 591–595, 2023.

[21] M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji, "SilentCipher: Deep audio watermarking," in *Interspeech 2024*. ISCA, Sep. 2024, p. 2235–2239.

[22] M. Moritz, T. Olán, and T. Virtanen, "Noise-to-mask ratio loss for deep neural network based audio watermarking," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, Sep. 2024, p. 1–6.

[23] C.-H. Huang and J.-L. Wu, "SLIC: Secure learned image codec through compressed domain watermarking to defend image manipulation," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, ser. MMAsia '24. New York, NY, USA: Association for Computing Machinery, 2024.

[24] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[25] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[26] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6309–6318.

[27] S. Sadok, J. Hauret, and É. Bavu, "Bringing interpretability to neural audio codecs," in *Interspeech 2025*. ISCA, Aug. 2025, p. 5023–5027.

[28] N. Tokui and T. Baker, "Latent granular resynthesis using neural audio codecs," 2025. [Online]. Available: https://arxiv.org/abs/2507.19202

[29] L. Juvela and X. Wang, "Audio codec augmentation for robust collaborative watermarking of speech synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2409.13382

[30] R. S. Roman, P. Fernandez, A. Deleforge, Y. Adi, and R. Serizel, "Latent watermarking of audio generative models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[31] Y. Liu, L. Lu, J. Jin, L. Sun, and A. Fanelli, "XAttn-Mark: Learning robust audio watermarking with cross-attention," in *Forty-second International Conference on Machine Learning*, 2025.

[32] A. Pujari and A. Rattani, "WaveVerify: A novel audio watermarking framework for media authentication and combatting deepfakes," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2025.

[33] L. Yao, C. Huang, S. Wang, J. Xue, H. Guo, J. Liu, P. Lin, T. Ohtsuki, and M. Pan, "Yours or mine? overwriting attacks against neural audio watermarking," 2025. [Online]. Available: https://arxiv.org/abs/2509.05835

[34] Y. Yao, J. Song, and J. Jin, "Hashed watermark as a filter: Defeating forging and overwriting attacks in weight-based neural network watermarking," 2025. [Online]. Available: https://arxiv.org/abs/2507.11137

[35] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 495–507, Nov. 2021.

[36] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[37] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech," 2025. [Online]. Available: https://arxiv.org/abs/2504.10782

[38] Y. Özer, W. Ge, Z. Zhang, X. Wang, and J. Yamagishi, "Self voice conversion as an attack against neural audio watermarking," 2026. [Online]. Available: https://arxiv.org/abs/2601.20432

[39] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, p. 2523–2533, Jun. 2023.

[40] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating music from text," 2023. [Online]. Available: https://arxiv.org/abs/2301.11325

[41] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[42] S. Chen, S. LIU, L. Zhou, E. Liu, X. Tan, J. Li, sheng zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," 2025.

[43] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: text-guided multilingual universal speech generation at scale," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[44] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "AudioPaLM: A large language model that can speak and listen," 2023. [Online]. Available: https://arxiv.org/abs/2306.12925

[45] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12 442–12 462.

[46] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International Conference on Digital Signal Processing (DSP)*, IEEE, IET, EURASIP. Santorini, Greece: IEEE, Jul. 2009, pp. 1–4.

[47] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[48] G. J. Mysore, "DAPS (device and produced speech) dataset," May 2014.

[49] Y. Özer, S. Schwär, V. Arifi-Müller, J. Lawrence, E. Sen, and M. Müller, "Piano Concerto Dataset (PCD): A multitrack dataset of piano concertos," *Trans. of the Int. Soc. for Music Inf. Retriev. (TISMIR)*, vol. 6, no. 1, pp. 75–88, 2023.

[50] T. Nakamura, S. Takamichi, N. Tanji, S. Fukayama, and H. Saruwatari, "jaCappella corpus: A Japanese a cappella vocal ensemble corpus," in *Proc. of the IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP)*, 2023.

[51] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019.

[52] Q. Xi, R. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "GuitarSet: A dataset for guitar transcription." in *Proc. of the Int. Soc. for Music Inf. Retriev. Conf. (ISMIR)*, Paris, France, 2018, pp. 453–460.

[53] T. Prätzlich, M. Müller, B. W. Bohl, and J. Veit, "Freischütz Digital: Demos of audio-related contributions," in *Demos and Late Breaking News of the Int. Soc. for Music Inf. Retriev. Conf. (ISMIR)*, Málaga, Spain, 2015.

[54] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time Speech Separation in the Time Domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6470–6474.

[55] T. Saeki, D. Xin, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.