GenMark: An Embedded Watermarking Scheme for Generative Audio Synthesis

Anonymous EMNLP submission

Abstract

Audio watermarking provides an effective approach for tracing and protecting synthetic audio content. Traditional methods often apply watermarking as a post-processing step, which makes the watermark vulnerable to removal or degradation through signal processing or model editing. To address these issues, our paper introduces GenMark, a novel approach that embeds watermarks directly into the decoder of neural audio generation models during training. Our approach combines time-frequency perceptual losses, a mask-based localization model, and adversarial training to ensure high audio quality and watermark robustness. Experimental results on speech and music generation tasks demonstrate superior detection accuracy (TPR: 99.9% for speech, 100.0% for music). GenMark also preserves perceptual quality with less than 2% degradation in MUSHRA scores, establishing it as a strong candidate for practical and secure watermarking in generative audio systems. The replication package can be accessed at the anonymous link.[1]

1 Introduction

With the rapid advancement and increasing accessibility of generative audio technologies (Xiang et al., 2017; You et al., 2021; Wang et al., 2023; Borsos et al., 2023; Suno, 2023; Copet et al., 2024), concerns about the abuse of synthetic speech are growing. Modern speech synthesis models like deepfake technology (Shaaban et al., 2023) enable voice cloning that could manipulate public discourse (News, 2024), damage individual reputations (Findlay, 2025), or compromise national security (Canadian Security Intelligence Service, 2023). These risks highlight the critical need for effective detection tools and traceability measures to verify the authenticity of synthetic audio and enforce accountability.

In such cases, audio watermarking serves as an effective solution by embedding imperceptible identifiers to trace model-generated audio, which effectively prevents malicious users' misuse of synthetic audio. Current mainstream audio watermarking methods embed watermarks directly into audio signals. In audio generation scenarios, it requires first generating audio by a generation model and then embedding a watermark into the generated audio. However, the post-processing watermarking strategy poses a serious security risk: malicious users can take control of the watermarking embedding process. By circumventing the watermark embedding stage, they are able to produce unwatermarked audio and exploit it in illicit scenarios. This poses a huge challenge to the regulation of synthetic audio. Moreover, audio generation poses unique challenges for watermarking, such as dealing with intricate frequency patterns and ensuring that the watermark stays reliable without affecting audio quality. These difficulties make it hard to embed robust and imperceptible watermarks.

To address these issues, we propose GenMark, a novel in-process injection watermark method that embeds the watermark during the audio generation process. GenMark improves traditional post-generation watermarking by directly generating audio with embedded watermarks. Unlike traditional post-generation watermarking methods, GenMark allows direct generation of audio with embedded watermarks. This prevents malicious attackers from manipulating the watermarking process and ensures reliable regulation of synthetic audio. Instead of modifying the entire generation pipeline, we focus only on the decoder, which converts tokens into audio samples. This choice enables efficient integration while maintaining generation quality. GenMark leverages joint time-frequency losses to improve perceptual audio quality and incorporates a mask model to enhance watermark robustness and location accuracy. In addition, it adopts

[1] https://anonymous.4open.science/r/Gen-Mark-1F27

GAN-based training to enhance the imperceptibility of the watermark. As a result, the generated waveforms inherently encode persistent and identifiable watermark signatures, regardless of input prompts or decoding parameters.

We evaluate GenMark using four state-of-the-art watermarking models, WavMark (Chen et al., 2024), AudioSeal (San Roman et al., 2024),SilentCipher (Singh et al., 2024), and Timbre (Liu et al., 2023a) on both speech and music generation tasks. In terms of audio quality, GenMark consistently achieves lower Frechet Audio Distance (FAD) and Kullback-Leibler Divergence (KLD) scores across multiple datasets, indicating minimal perceptual and distributional distortion. It also maintains strong semantic alignment, outperforming baselines on the CLAP metric. For detection, we report **TPR**, **FPR**, and **decode accuracy**. Our method outperforms baselines in both detection and watermark recovery. To evaluate robustness, we subject watermarked audio to 11 common audio transformations and adversarial attacks, comparing the decoding error rates with those of WavMark, AudioSeal, and SilentCipher. Besides, subjective MUSHRA evaluations further confirm that GenMark preserves perceptual quality and the ablation studies show that each component of GenMark contributes to the balance between fidelity, robustness, and detection precision. We summarize contributions as follows:

- We propose GenMark, a novel framework that embeds inaudible watermarks directly into generative audio models during training.
- GenMark introduces a multi-scale discriminator and a mask model to improve audio quality and watermark robustness.
- Experiments show near-perfect detection rates (TPR: 99.9% for *Bark*, 100.0% for MusicGen) with FPR $\leq$0.1%. GenMark maintains low decoding error rates under 11 distortions and less than 2% perceptual degradation in MUSHRA tests, outperforming state-of-the-art baselines.

2 Preliminaries

2.1 Audio Generation

The current neural audio generation systems follow a hierarchical processing pipeline. Multi-modal inputs—such as text or speech prompts—are first encoded into discrete acoustic tokens through cascaded transformer layers (Vaswani et al., 2017). These tokens serve as high-level latent representations of the target audio. To synthesize natural-sounding waveforms, the tokens are then passed through spectral enhancement modules, including neural vocoders (Kong et al., 2020) and differentiable signal processing components (Engel et al., 2020). Finally, the decoder transforms the processed acoustic tokens and synthesizes them into the final audio waves.

2.2 Loss Balancer

In multi-objective training settings, gradients from different loss terms can vary significantly in scale. This imbalance may lead to unstable optimization and make the effect of each loss weight λ hard to interpret. To address this, we adopt loss balancers inspired by EnCodec (Défossez et al., 2022), which dynamically rescales gradient contributions based on their recent magnitude.

For each loss $\mathcal{L}_i$, we compute its gradient $g_i = \frac{\partial \mathcal{L}}{\partial \hat{x}}$ and track the exponential moving average of its norm $\|g_i\|_2^\beta$. Then, the rescaled gradient is,

$$\tilde{g}_i = \frac{R \cdot \lambda_i}{\sum_j \lambda_j} \cdot \frac{g_i}{\|g_i\|_2^\beta}. \quad (1)$$

The final gradient used for backpropagation is $\sum_i \tilde{g}_i$, instead of the original $\sum_i \lambda_i g_i$, which helps stabilize training. The R is a reference gradient scale, and the β is a decay rate.

3 Methodology

GenMark employs gradient steganography to embed watermark signals directly into the generative process by optimizing the decoder component of the model. Instead of modifying the entire generation pipeline—which is often large and difficult to fine-tune—we target the decoder, the final stage responsible for converting discrete token sequences into audio waveforms. This position makes it particularly suitable for learning robust watermark patterns. By training the decoder to produce watermarked audio without compromising perceptual quality, we enable direct integration of watermarking into the model. Once trained, the decoder can be seamlessly substituted for the original one, enabling watermark embedding without modifying the rest of the generation pipeline.

3.1 Train Pipeline

Overview. To embed watermark information m into the parameters of the decoder, we guide the decoder's optimization using a joint loss, which includes perceptual loss ($\mathcal{L}_{\text{time}}, \mathcal{L}_{\text{spec}}$), adversarial loss ($\mathcal{L}_{\text{gen}}, \mathcal{L}_{\text{disc}}, \mathcal{L}_{\text{feat}}$), and decoding loss ($\mathcal{L}_{\text{msg}}$).

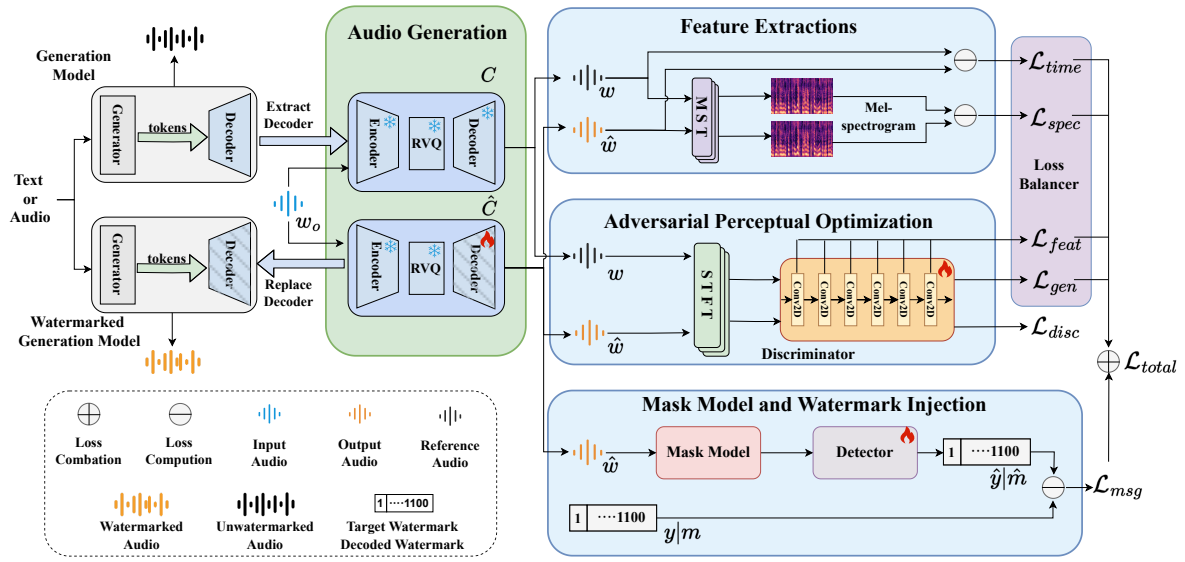

Figure 1: Overview of the training pipeline. We use two audio codecs in this framework: C is a frozen reference codec used to produce clean (unwatermarked) audio, while $\hat{C}$ is a trainable version where only the decoder is updated to embed watermarks. Losses are computed between clean and watermarked outputs to maintain quality. A mask model further improves robustness, and a decoder network extracts the watermark from the output. In addition to standard perceptual and adversarial losses, $\mathcal{L}_{\text{dic}}$is used to optimize the discriminator, and the weight of the message loss $\mathcal{L}_{\text{msg}}$is tuned separately to ensure effective watermark embedding.

We balance these objectives during training by scaling their gradient contributions using the Loss Balancer 2.2. The full training pipeline consists of four stages, as illustrated in Figure 1.

Audio generation. Firstly, we extract the compression model (e.g., EnCodec) from the audio generation model (e.g., *Bark*). The codec $\hat{C}$ consists of an encoder, a quantizer, and a decoder, which together map raw waveforms to discrete tokens and reconstruct audio from them. During optimization, we freeze the encoder and quantizer of $\hat{C}$ and only update its decoder, which converts tokens into waveforms. This setup enables efficient watermark embedding by modifying only the decoder.

Given an input audio signal $w_o \in \mathbb{R}^T$, the codec $\hat{C}$ generates a watermarked version $\hat{w} \in \mathbb{R}^T$. For reference, we use an untrained copy of the same codec, denoted C, to reconstruct a non-watermarked version w. In the subsequent training steps, we simultaneously optimize for two objectives: enabling reliable watermark decoding from $\hat{w}$, and minimizing the difference between w and $\hat{w}$ to preserve audio quality.

Feature extractions. To preserve perceptual audio quality, we compute time-domain $\mathcal{L}_{\text{time}}$and frequency-domain $\mathcal{L}_{\text{spec}}$losses between w and $\hat{w}$. The time-domain loss constrains waveform-level distortions, promoting time-domain alignment. The frequency-domain loss is calculated using multi-scale Mel spectrograms, which are widely used to reflect human auditory perception and capture perceptual differences across resolutions (Kong et al., 2020; You et al., 2021). This hybrid loss strategy has been shown effective in maintaining perceptual fidelity in neural audio synthesis (Tan et al., 2024; Zhang et al., 2019; Yamamoto et al., 2020) and compression tasks (Défossez et al., 2022; Zeghidour et al., 2021).

Adversarial Perceptual Optimization. To improve audio quality and reduce perceptual artifacts, we adopt adversarial training following prior works (Défossez et al., 2022). As illustrated in Figure 1, the decoder of the codec serves as the generator, producing watermarked audio $\hat{w}$, while a lightweight multi-scale discriminator distinguishes $\hat{w}$ from the reference audio w. The adversarial loss for the generator is $\mathcal{L}_{\text{gen}}$and for the discriminator is $\mathcal{L}_{\text{dic}}$. Similarly to previous work (You et al., 2021; Kong et al., 2020), we also incorporate a feature-matching loss $\mathcal{L}_{\text{feat}}$for the generator.

Maks Model and Watermark Injection. The watermarked audio $\hat{w}$ is further processed by a mask model M (in Section 3.4) designed to enhance robustness and enable fine-grained watermark localization. The model comprises two components: a Localization Refinement Module, and a Robustness Enhancement Module module. They ensures the watermark remains detectable under common

audio modifications while reducing false positives. After that, the audio is fed to the watermark detector D_{det}, which outputs $D_{det}(\hat{w}) \in [0,1]^{18\times T}$. The first two dimensions of $D_{det}(\hat{w})$ represent the frame-level probabilities of watermark presence, while the remaining 16 dimensions correspond to the decoded 16-bit watermark sequence. This prediction is then compared with the target watermark message m, and the discrepancy is used to compute the decoding loss $\mathcal{L}_{\text{msg}}$, guiding the model to embed the watermark into the audio. The architecture details of D_{det} are provided in the Appendix E.

3.2 Feature Extractions.

Although the primary objective is to embed watermark signals into the audio, it is crucial that the perceptual quality of the output remains unaffected. To ensure this, the audio fidelity loss incorporates complementary constraints across both time and frequency domains, informed by principles of human auditory perception (Xiang et al., 2017).

$$\mathcal{L}_{\text{time}} = \|w - \hat{w}\|_1, \tag{2}$$

Eq. (2) promotes robust waveform similarity while remaining minor phase variations that have minimal perceptual impact (Engel et al., 2020).

However, as human auditory perception varies in sensitivity across different frequency ranges, optimization in the time domain alone may not suffice to achieve high-quality audio perception. To address this, we introduce a Multi-scale Mel Spectrogram Loss (Gritsenko et al., 2020), which constrains the spectral characteristics (frequency domain feature) of the generated audio. Eq. (3) uses a multi-resolution Mel-spectrogram analysis with window sizes set $\mathcal{H} = \{32, 64, 128, 256, 512, 1024\}$. And $S_h(\cdot)$ denotes the function of the Mel-spectrogram using a fixed window size h:

$$\mathcal{L}_{\text{spec}} = \sum_{h\in\mathcal{H}} \sum_{i=1,2} \left[\|S_h(w) - S_h(\hat{w})\|_i \right]. \tag{3}$$

The combination of absolute difference (ℓ_1) and squared difference (ℓ_2) formulation balances spectral magnitude alignment with overall distribution consistency (Gritsenko et al., 2020), reducing the over-smoothing effects often observed in pure ℓ_2 optimization (Kong et al., 2020).

3.3 Adversarial Perceptual Optimization.

Although feature-based losses help maintain the overall perceptual quality of audio, they may not fully capture subtle distortions or unnatural details that can still affect the quality of audio. To further enhance perceptual realism and improve watermark robustness, we adopt an adversarial training strategy using multi-scale spectral discriminators, inspired by prior work on neural vocoders and audio synthesis (Défossez et al., 2022; You et al., 2021; Kong et al., 2020).

The discriminator architecture follows a five-layer dilated convolutional design with dilation rates $[1, 2, 4]$, weight normalization, and LeakyReLU activations ($\alpha = 0.2$) for stable convergence. It processes the input across multiple spectral resolutions in parallel, using STFTs with FFT sizes $\{512, 1024, 2048\}$ and corresponding window lengths $\{128, 256, 512\}$. This multi-scale structure enables the discriminator to capture both fine- and coarse-grained spectral artifacts, making it a strong perceptual sensitivity.

Generator Objective. The generator G is trained to generate watermarked audio that is perceptually indistinguishable from original signals:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\hat{w}} \left[\mathbb{E}_{k\in\mathcal{K}} \|1 - D_k(\hat{w})\|_1 \right], \tag{4}$$

where $\mathcal{K}$ represents the STFT window size set set, and$D_k(\cdot)$ is the discriminator output.

In addition, inspired by prior work (Kumar et al., 2019a; Kong et al., 2020; You et al., 2021; Défossez et al., 2022), we include a feature-matching loss $\mathcal{L}_{\text{feat}}$encourages the generator to produce internal representations that closely resemble those extracted from real audio by the discriminator:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{l\in\mathcal{S},k\in\mathcal{K}} \left[\frac{\|D_k^l(w) - D_k^l(\hat{w})\|_1}{\mathbb{E}[D_k^l(w)] + \epsilon} \right], \tag{5}$$

where $\mathcal{S}$ denotes the set of discriminator layers, and D_k^l represents the output of the l-th layer of the discriminator corresponding to an STFT window size k. The term $\epsilon = 10^{-6}$ is introduced to prevent division by zero.

Discriminator Objective. The discriminator D is optimized to differentiate between real and watermarked audio signals:

$$\begin{aligned}\mathcal{L}_{\text{dic}} = \mathbb{E}_{w} \left[\mathbb{E}_{k\in\mathcal{K}} \|1 - D_k(w)\|_1 \right] \\ + \mathbb{E}_{\hat{w}} \left[\mathbb{E}_{k\in\mathcal{K}} \|D_k(\hat{w})\|_1 \right].\end{aligned} \tag{6}$$

3.4 Maks Model and Watermark Injection

3.4.1 Maks Model

In order to reduce the false positive rate, improve localization accuracy, and enhance watermark robustness, we additionally include an enhanced mask

module, which exposes the decoder to a variety of masking patterns during training, enabling it to better distinguish true watermark signals, improve its resilience to common audio attacks.

(1) Localization Refinement Module: To reduce false positives and improve spatial precision, this module introduces two training strategies: (a) part of watermarked segments are replaced with alternative watermark patterns to prevent overfitting and improve generalization; (b) within each audio, K regions are randomly selected and partially replaced with clean, unrelated, or silent content. These perturbations force the decoder to learn precise localization and improve extraction accuracy by distinguishing true watermark regions from distractors. The detailed parameter settings are provided in Appendix B.

(2) Robustness Enhancement Module: To improve the watermark's resilience to signal processing attacks, we develop a sequential transformation pipeline that applies nine fundamental audio operations in carefully calibrated proportions, including frequency filtering, resampling, dynamic range adjustment, echo effects, noise addition, and waveform smoothing. This transformation is commonly used in watermark removal attacks and watermark robustness enhancement (Kirovski and Malvar, 2003; Li et al., 2024). By simulating these attacks during training, the decoder learns to maintain watermark fidelity. The probability and parameters of each operation (e.g., frequency thresholds for filtering, signal strength for noise addition) are carefully optimized, as outlined in Appendix C.

3.4.2 Watermark Injection

To ensure stable and accurate watermark recovery, we define a message loss that guides the model to retain the correct message content during decoding. It consists of two core components:

$$\begin{cases} \mathcal{L}_{\text{det}} = \frac{1}{T}\sum_{t=1}^{T}\left[\text{BCE}\left(y_t, \hat{y}_t\right)\right] \\ \mathcal{L}_{\text{payload}} = \frac{1}{T}\sum_{t=1}^{T}\left[\text{BCE}\left(m_t, \hat{m}_t\right)\right], \end{cases} \tag{7}$$

where $y_t \in \{0,1\}$ denotes the presence of a watermark in frame t, and $m_t \in \{0,1\}^{16}$ corresponds to the ground-truth 16-bit message. The overall watermark loss function is formulated as:

$$\mathcal{L}_{\text{msg}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{payload}}\mathcal{L}_{\text{payload}}, \tag{8}$$

where λ_{det} and λ_{payload} balance the importance of detection accuracy and payload reconstruction. As described in the training pipeline, the decoder D receives the masked audio output from the Mask Model and produces a tensor $D(\hat{w}) \in [0,1]^{18\times T}$, where each of the T frames contains detection and decoding information. Guided by $\mathcal{L}_{\text{msg}}$, the decoder is trained not only to accurately determine which frames contain watermark content, but also to maintain robustness against typical audio attacks. This enables precise frame-level localization of embedded watermarks and ensures reliable decoding performance even under signal distortions.

4 Experiments Setting

Models and Datasets. We use two state-of-the-art generative models, *Bark* (Suno, 2023) for speech synthesis and *MusicGen* (Copet et al., 2024) for musical audio generation, to insert watermark. Training and evaluation are conducted on `AudioSet` (Gemmeke et al., 2017) and `CommonVoice` (Foundation, 2020) dataset, ensuring diverse coverage of both general acoustic environments and multilingual speech. Since *Bark* requires textual prompts as input, we additionally incorporate several text-based datasets as test cases to evaluate watermarking performance: HarvardSentences (on Subjective Measurements, 1969), LibriSpeech (Panayotov et al., 2015) and LJSpeech (Ito, 2017). These setups enable a comprehensive assessment of our watermarking method across speech and non-speech domains.

Training Configuration. All models are trained on an NVIDIA RTX 3090 GPU with an initial learning rate of 1×10^{-4}, which is gradually decreased for stable convergence. Batch sizes are set to 24 for *Bark* and 16 for MusicGen, reflecting their respective computational demands. To accommodate the inherent sampling preferences of these models, *Bark* is trained at 24 kHz, while MusicGen is trained at 32 kHz. We balance our multi-objective loss using the balancer with $\lambda_{\text{time}} = 1$, $\lambda_{\text{freq}} = 6$, $\lambda_{\text{gen}} = 9$, $\lambda_{\text{feat}} = 9$, $\lambda_{\text{msg}} = 10$. The discriminator updates once every two epochs, allowing the generator sufficient adaptation time and ensuring more stable adversarial training dynamics.

Baselines. Our method is benchmarked against several competitive baselines: AudioSeal (San Roman et al., 2024), Wavmark (Chen et al., 2024), SilentCipher (Singh et al., 2024), and Timbre (Liu et al., 2023a). These methods are recognized for their effectiveness in audio watermarking, and together, they provide a strong benchmark for evaluating im-

Dataset	LibriSpeech			HarvardSentence			LJSpeech		
Model	KLD	CLAP	FAD	KLD	CLAP	FAD	KLD	CLAP	FAD
AudioSeal	0.3360	13.90	0.6727	0.2029	9.28	0.4533	0.1727	8.67	0.1976
WavMark	0.3926	13.89	1.3210	0.1526	9.15	1.6092	0.1641	9.49	1.5716
SilentCipher	0.3242	14.21	0.3251	0.1370	9.39	0.2936	0.1375	8.81	0.1794
Timbre	0.3961	14.00	0.9855	0.1704	9.29	0.7345	0.1588	8.57	0.5186
Ours (GenMark)	0.3234	13.86	0.0957	0.1321	9.28	0.0615	0.1364	8.11	0.0227

Table 1: Model Comparison under Perceptual / Distributional Metrics. We use publicly available implementations for CLAP and FAD

perceptibility, robustness, and decoding accuracy across a range of audio conditions.

5 Experiments Result

We evaluate GenMark in four key aspects: audio quality, detection accuracy, robustness, and human perception. Specifically, we assess whether watermarking affects perceptual and semantic quality, measure detection performance across different models, test robustness under common audio perturbations, and conduct a subjective listening study to understand the impact on human listeners. In addition, we conduct ablation studies to validate the effectiveness of key components, including the mask model and the discriminator, in improving watermark imperceptibility and robustness.

5.1 Quality of Audio

To explore GenMark's capability to preserve perceptual and semantic quality in synthetic audio, we evaluate the similarity between the generated watermarked audio and original audio samples based on perceptual, distributional, and semantic metrics.
Perceptual and Distributional Quality. We use the Frechet Audio Distance [2] (FAD) (Kilgour et al., 2018), a reference-free perceptual quality metric adapted from the Frechet Inception Distance (FID). In addition, we use Kullback-Leibler Divergence (KLD) to measure the distributional deviation between original and watermarked audio.
Table 1 shows that GenMark consistently achieves lower FAD scores across all tested datasets. Specifically, GenMark achieves significantly lower FAD scores (0.0957 for LibriSpeech, 0.0615 for HarvardSentence, and 0.0227 for LJSpeech) compared to existing methods, indicating minimal perceptual distortion. Furthermore, GenMark also achieves the lowest KLD values across all datasets (0.3234 for LibriSpeech, 0.1321 for HarvardSentence, and 0.1364 for LJSpeech), signifying excellent preservation of distributional characteristics.

[2] https://github.com/microsoft/fadtk

Model	*Bark*			*Musicgan*		
	TPR	FPR	Acc	TPR	FPR	Acc
Audioseal	100.0	0.0	95.4	100.0	0.1	73.3
Wavmark	99.8	0.0	99.8	95.2	0.1	94.4
SilentCipher	92.4	31.4	96.6	98.2	39.6	97.8
Timbre	\	\	99.9	\	\	94.7
Ours	99.9	0.0	99.8	100.0	0.1	94.3

Table 2: Detection results for *Bark*, *Musicgan* with TPR, FPR and Decode Accurate (%).

Semantic Consistency. We use the CLAP [3] metric derived from a contrastive language-audio pretraining model to assess semantic preservation, which reflects alignment between the generated audio and the original text prompt. Lower CLAP values imply better semantic retention. GenMark achieves the best CLAP scores on LibriSpeech (13.86) and LJSpeech (8.11), outperforming all baselines. On HarvardSentence, it is slightly behind WavMark (9.28 vs. 9.15), but still ahead of other methods. These results demonstrate that GenMark consistently preserves semantic alignment while embedding watermark signals.

5.2 Detection Accuracy

To assess the efficacy of our watermarking technique, we conducted comprehensive detection experiments using two prominent generative audio models: *Bark* and *MusicGen*. *Bark* is designed for high-quality speech synthesis, whereas *MusicGen* is tailored for generating musical audio. This selection enables a robust evaluation of the watermarking technique across both linguistic and musical contexts. We generate and analyze 10,000 audio samples per method for each model to ensure statistically reliable results. Detection performance is measured using TPR and FPR, as presented in Table 2. TPR quantifies the proportion of correctly identified watermarked samples, while FPR reflects the ratio of non-watermarked samples being incorrectly flagged (Gong et al., 2024a).
As shown in Table 2, GenMark has strong detec-

[3] https://github.com/LAION-AI/CLAP

Model	Audio Transformations (Decoding Error Rates %)											Total
	Bandpass	Highpass	Lowpass	Speed	Resample	Boost	Duck	Echo	Pink	White	Smooth	
AudioSeal	92.08	100.00	100.00	99.85	4.66	29.24	95.63	15.61	23.68	24.65	16.53	54.72
WavMark	**0.21**	**0.13**	100.00	97.57	**0.12**	7.83	4.89	3.95	79.23	99.72	26.33	38.27
SilentCipher	34.58	43.26	97.70	99.16	7.01	6.66	6.78	79.79	100.00	100.00	70.06	58.72
Ours (GenMark)	**0.05**	68.57	**50.97**	**1.36**	0.15	**1.24**	**0.17**	**0.17**	**1.62**	**3.12**	**1.04**	**11.55**

Table 3: Decoding Error Rates (%) under different audio transformations.

tion performance. For the *Bark* model, our method achieves a TPR of 99.9% with zero false positives, while attaining perfect detection (100.0% TPR) on *MusicGen* with a minimal FPR of 0.1%. These results highlight the precision and robustness of our detector. Although AudioSeal also achieves high TPRs, especially on *Bark*, it shows a noticeable drop in accuracy on *MusicGen*. In contrast, our method maintains balanced performance across both domains. WavMark exhibits similar accuracy to our method on *Bark* but falls short in TPR on *MusicGen*. SilentCipher's performance is less stable overall, with high false positives observed in both settings. Besides, Timbre does not support watermark detection, and only supports watermark decoding. As such, TPR and FPR are not applicable in this context. The consistent performance across diverse audio domains highlights GenMark's suitability for practical.

5.3 Robustness of Watermark

To evaluate the robustness of GenMark under real-world perturbations, we conduct a comprehensive benchmark using the *Bark* model as the generative backbone. We evaluate robustness by applying 11 common audio transformations, including various signal processing operations, dynamic range modifications, ambient noise interference, and smoothing. For each transformation, we calculate the decode error rate—the percentage of watermark decoded incorrectly. Lower values mean better robustness.

As shown in Table 3, GenMark achieves the lowest error rates in 9 out of 11 transformations. It handles distortions like echo, ducking, background noise, and speed change especially well, with error rates often below 2%. Even under challenging conditions like lowpass filtering, where most methods fail completely, our model reduces the error rate to around 51%.

Although performance drops with highpass filtering, GenMark still remains competitive overall. Its average decode error rate is just 11.55%, much lower than WavMark (38.27%), SilentCipher(58.72%), and AudioSeal (54.72%). This shows that our method keeps the watermark stable and decodable even after heavy audio processing.

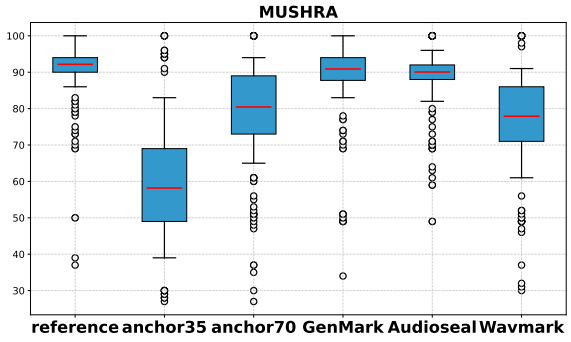

Figure 2: Distribution of MUSHRA scores for watermarked audio in the subjective evaluation study. Some extreme outliers beyond the whisker range are marked.

5.4 Usable Study

To assess perceptual audio quality from a human perception perspective, we perform a subjective evaluation using the standardized MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) protocol (ITU-T, 2015), a well-established methodology widely adopted for audio quality benchmarking. We invite 20 audio experts to evaluate 20 audio groups, each corresponding to a distinct prompt. For every prompt, one sample was randomly selected from 100 *Bark*-generated clips. Each group includes the following: (1) three types of watermarked audio samples (GenMark, AudioSeal, WavMark); (2) one clean reference; and (3) two anchor signals, namely Anchor35 (filtered at 3.5 kHz) and Anchor70 (filtered at 7 kHz). Participants rate each sample on a scale of 0–100, with anchors and references used to guide their judgments. Details are provided in Appendix F.

As presented in Figure 2, our proposed method attains the highest MUSHRA score (**90.89**), closely followed by AudioSeal (**90.06**), with WavMark lagging at **77.90**. These results demonstrate that both our method and AudioSeal effectively preserve perceptual audio quality, whereas WavMark introduces perceptible degradation. For comparison, the clean reference audio achieves a MUSHRA score of **92.17**, while the Anchor70 and Anchor35 con-

ditions score **80.44** and **58.18**, respectively. These results confirm the evaluators' consistency and sensitivity in the subjective assessment.

5.5 Ablation study

To understand the contribution of each component in our framework, we conduct an ablation study focusing on three core modules: (1) adversarial perceptual optimization, (2) the robustness enhancement module, and (3) the localization refinement module. For each variant, we remove or disable one of the components and evaluate performance on key metrics, including detection (TPR, FPR, Acc), perceptual quality (FAD), and robustness (average decode error rate under transformations), as shown in Table 4.

Removing adversarial training (*NoAdversarial*) results in a drop in perceptual quality, as indicated by the increase in FAD from 0.0615 (full model) to 0.1493. Disabling the robustness enhancement module (*NoRobustMask*) has the most significant effect on robustness, with the average decode error rate (DER) surging from 11.55% to 46.22%. Removing the localization refinement module (*NoLocMask*) improves robustness but at the cost of a substantial increase in FPR, highlighting its importance in maintaining detection precision.

Variant	TPR↑	FPR↓	Acc↑	FAD↓	DER↓
NoAdversarial	99.8	0.0	99.8	0.1493	10.92
NoRobustMask	99.9	0.0	99.9	0.0607	46.22
NoLocMask	100.0	35.7	100.0	0.0595	6.43
Full Model	**99.9**	**0.0**	**99.8**	**0.0615**	**11.55**

Table 4: Ablation study of GenMark evaluating the effect of each component on detection (TPR, FPR, Acc), perceptual quality (FAD), and robustness (DER).

6 Related Work

6.1 Audio Generation

Currently, audio generation has evolved significantly through deep learning. For instance, autoregressive models such as WaveNet (Van Den Oord et al., 2016) greatly improve audio quality, whereas GAN-based approaches, like MelGAN (Kumar et al., 2019b), enhance synthesis efficiency. These advancements established the foundation for contemporary neural audio generation techniques.

Recent studies integrate transformers and diffusion models to achieve further development for audio generation. AudioLDM (Liu et al., 2023b) uses contrastive language audio pretraining (Wu et al., 2023) with latent diffusion (Rombach et al., 2022) for text-guided generation. Audio language models such as *Bark* (Suno, 2023), MusicGAN (Copet et al., 2024), and AudioLM (Borsos et al., 2023) use text-generation techniques (Radford, 2018; Brown et al., 2020a), encoding text and timbre into tokens using EnCodec (Défossez et al., 2022) and SoundStream (Zeghidour et al., 2021) for transformer-based sequence-to-sequence synthesis.

6.2 Audio Watermark

Traditional audio watermarking techniques (Cvejic and Seppanen, 2004; Anderson, 1996) typically embed watermarks by manipulating information in the time or frequency domains (Cox et al., 1997; Xiang et al., 2018; Su et al., 2018; Liu et al., 2019). These methods depend on manually crafted heuristic rules and specialized domain expertise to guide their design and implementation. Simultaneously achieving a high imperceptibility, capacity, and robustness watermark across diverse audio types remains a significant challenge.

With advancements in deep learning, the ability to automatically learn watermark embedding and extraction techniques has simplified the design of watermarking methods (Tai and Mansour, 2019; Pavlović et al., 2022). In particular, current deep learning-based watermarking techniques generally follow an Encoder-Decoder structure (Qu et al., 2023; Ren et al., 2023; Chen et al., 2024; San Roman et al., 2024), where the encoder generates watermarked audio, and the decoder extracts the information from the watermarked audio. The entire model is trained in an end-to-end manner, enabling it to automatically learn the wa termark embedding and extraction processes.

7 Conclusion

This work introduces GenMark, a robust and efficient method for embedding traceable, imperceptible watermarks directly into generative audio models. By integrating watermark objectives directly into the generation model, GenMark effectively addresses the vulnerabilities of traditional post-generation watermarking. Extensive evaluation across speech and music generation domains confirms that GenMark offers superior detection accuracy, resilience to a wide array of audio attacks, and negligible perceptual degradation. These results establish GenMark as a strong tool for safeguarding audio synthesis systems.

Limitations

While GenMark demonstrates strong performance across multiple generative audio tasks, it requires model-specific integration during training. Since the watermark is embedded directly into the decoder, each generative model (e.g., Bark, MusicGen) must be individually fine-tuned with GenMark.

References

2020. Fraudsters use ai-generated voices to scam companies. *The Wall Street Journal*. https://www.wsj.com.

Ross Anderson. 1996. Information hiding: First international workshop cambridge, uk, may 30–june 1, 1996 proceedings. In *International Workshop on Information Hiding 1*. Springer.

Mark Arnold, Martin Schmucker, and Stephen D Wolthusen. 2003. Techniques and applications of digital watermarking and content protection. *Artech House*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Canadian Security Intelligence Service. 2023. The evolution of disinformation: A deepfake future. Accessed: 2025-05-15.

Nicholas Carlini et al. 2023. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.10917*.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. 2024. Wavmark: Watermarking for audio generation. *Preprint*, arXiv:2308.12770.

Xing Chen, Han Liu, et al. 2023. Wavmark: Imperceptible and robust audio watermarking. In *ICLR*.

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

I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. 1997. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687.

Ingemar J Cox, Matthew L Miller, and Jeffrey A Bloom. 2007. *Digital watermarking and steganography*. Morgan Kaufmann.

N. Cvejic and T. Seppanen. 2004. Increasing robustness of lsb audio steganography using a novel embedding method. In *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, volume 2, pages 533–537 Vol.2.

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

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. 2020. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

Georgina Findlay. 2025. 'you're gonna find this creepy': my ai-cloned voice was used by the far right. could i stop it? *Tech News Wires*. Accessed: 2025-05-15.

Mozilla Foundation. 2020. Common voice: A massively multilingual speech corpus.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Chen Gong, Kecen Li, Jin Yao, and Tianhao Wang. 2024a. Trajdeleter: Enabling trajectory forgetting in offline reinforcement learning agents. *arXiv preprint arXiv:2404.12530*.

Chen Gong, Zhou Yang, Yunpeng Bai, Jieke Shi, Junda He, Kecen Li, Bowen Xu, Sinha Arunesh, Xinwen Hou, David Lo, and Tianhao Wang. 2024b. Baffle: Hiding backdoors in offline reinforcement learning datasets. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 218–218. IEEE.

Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, and Nal Kalchbrenner. 2020. A spectral energy distance for parallel speech synthesis. *Advances in Neural Information Processing Systems*, 33:13062–13072.

Keith Ito. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

ITU-T. 2015. Method for the subjective assessment of intermediate quality levels of coding systems. Recommendation ITU-R BS.1534-3, International Telecommunication Union, Geneva, Switzerland. Also known as MUSHRA (Multiple Stimuli with Hidden Reference and Anchor).

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Darko Kirovski and Henrique S Malvar. 2003. Spread-spectrum watermarking of audio signals. *IEEE transactions on signal processing*, 51(4):1020–1033.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019a. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019b. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Pengcheng Li, Xulong Zhang, Jing Xiao, and Jianzong Wang. 2024. Ideaw: Robust neural audio watermarking with invertible dual-embedding. *arXiv preprint arXiv:2409.19627*.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023b. Audioldm: Text-to-audio generation with latent diffusion models. *Preprint*, arXiv:2301.12503.

Zhenghui Liu, Yuankun Huang, and Jiwu Huang. 2019. Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks. *IEEE Transactions on Information Forensics and Security*, 14(5):1171–1180.

Chengcheng Mou, Zhiyao Zhang, et al. 2023. Audioseal: Audio watermarking as a defense against speech deepfakes. In *NeurIPS*.

ABC News. 2024. Fake biden robocall urges new hampshire voters to skip their primary. Accessed: 2025-05-15.

IEEE Subcommittee on Subjective Measurements. 1969. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

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

Xinghua Qu, Xiang Yin, Pengfei Wei, Lu Lu, and Zejun Ma. 2023. Audioqr: Deep neural audio watermarks for qr code. In *IJCAI*, pages 6192–6200.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Yanzhen Ren, Hongcheng Zhu, Liming Zhai, Zongkun Sun, Rubing Shen, and Lina Wang. 2023. Who is speaking actually? robust and versatile speaker traceability for voice conversion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8674–8685.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

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

Ousama A. Shaaban, Remzi Yildirim, and Abubaker A. Alguttar. 2023. Audio deepfake approaches. *IEEE Access*, 11:132652–132682.

Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji. 2024. Silentcipher: Deep audio watermarking. *arXiv preprint arXiv*, 2406:03822.

Zhaopin Su, Guofu Zhang, Feng Yue, Lejie Chang, Jianguo Jiang, and Xin Yao. 2018. Snr-constrained heuristics for optimizing the scaling parameter of robust audio watermarking. *IEEE Transactions on Multimedia*, 20(10):2631–2644.

Inc Suno. 2023. Bark: A text-to-audio model. `https://github.com/suno-ai/bark`. MIT License.

Yuan-Yen Tai and Mohamed F Mansour. 2019. Audio watermarking over the air with modulated self-correlation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2452–2456. IEEE.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yipeng Wang, Min Wu, Wade Trappe, and KJ Ray Liu. 2004. Attacks on digital audio watermarks: Classification, evaluation benchmarks, and countermeasures. *Proceedings of the IEEE*, 92(6):921–936.

Xiaojun Wu, Tianyu Lin, and Wei Zhang. 2022. Adversarial attacks on neural watermarking models. *ICASSP*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yong Xiang, Guang Hua, Bin Yan, Yong Xiang, Guang Hua, and Bin Yan. 2017. Human auditory system and perceptual quality measurement. *Digital Audio Watermarking: Fundamentals, Techniques and Challenges*, pages 7–27.

Yong Xiang, Iynkaran Natgunanathan, Dezhong Peng, Guang Hua, and Bo Liu. 2018. Spread spectrum audio watermarking using multiple orthogonal pn sequences and variable embedding strengths and polarities. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):529–539.

Meng Xu et al. 2023. Evading detection: Towards robust steganography against diffusion-based detectors. *arXiv preprint arXiv:2310.01247*.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE.

Jaeseong You, Dalhyun Kim, Gyuhyeon Nam, Geumbyeol Hwang, and Gyeongsu Chae. 2021. Gan vocoder: Multi-resolution discriminator is all you need. *arXiv preprint arXiv:2103.05236*.

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

Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. 2019. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):631–644.

A Extended related work

A.1 Security and Misuse in Generative Models

The rapid advancement of generative models across text (Brown et al., 2020b; Touvron et al., 2023), image (Rombach et al., 2022), and audio (Kreuk et al., 2022) domains has brought remarkable synthesis quality and expressiveness. However, with this growth comes increasing concern over misuse. Recent work has shown that generative pipelines can be tampered with or exploited, such as backdoor injection in offline reinforcement learning datasets (Gong et al., 2024b), data poisoning in large language models (Carlini et al., 2023), and output evasion in diffusion models (Xu et al., 2023). These studies highlight the importance of security-aware generative model design, especially in ensuring traceability and tamper resistance.

In the audio domain, the risk is amplified by the realism of synthetic speech. Voice cloning and TTS systems have been used for impersonation, misinformation (News, 2024), and fraud (wes, 2020). Watermarking has emerged as a defense strategy (Mou et al., 2023; Chen et al., 2023; Singh et al., 2024), yet most approaches apply watermarks after generation, leaving them vulnerable to removal or circumvention. Our work addresses

this gap by embedding watermarks directly during training, offering stronger protection against post-generation manipulation.

A.2 Compression Model

SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022) are neural audio codecs designed for high-fidelity audio compression and reconstruction. SoundStream introduces a fully learnable end-to-end framework using residual vector quantization, while EnCodec builds upon this design with improved scalability and audio quality through hierarchical quantization and adversarial training. These models pioneer neural audio compression through self-supervised learning and hierarchical quantization. Unlike traditional handcrafted feature methods, these approaches efficiently encode high-dimensional audio into discrete tokens, retaining semantic information.

This tokenization framework empowers Transformer-based systems (e.g., *Bark* (Suno, 2023), MusicGAN (Copet et al., 2024), AudioLM (Borsos et al., 2023)) to perform cross-modal audio generation from text prompts and context-aware audio continuation. By integrating audio compression with language model architectures, these methods improve efficiency and versatility in generative AI, facilitating a wide range of multimodal synthesis applications.

A.3 Attacks on Audio Watermarking Systems

While audio watermarking enables traceability of generated content, ensuring robustness under adversarial or lossy conditions remains a major challenge. Watermarks are often vulnerable to signal manipulations such as compression, noise injection, cropping, pitch shifting, or time-stretching (Cox et al., 2007; Arnold et al., 2003). Attackers can intentionally apply these distortions to remove or degrade the watermark information without significantly affecting audio perceptual quality.

Classical attack strategies include re-encoding, filtering, jittering, or frequency band removal (Wang et al., 2004). Recent works even explore adversarial perturbations designed specifically to confuse watermark extractors (Wu et al., 2022). Therefore, the evaluation of watermark robustness must consider both standard degradations (e.g., MP3 compression, resampling) and targeted attacks (e.g., masking, inversion, audio remix).

In our experiments, we systematically test `GenMark` under 11 widely used audio transformations and adversarial manipulations to benchmark its resistance. Our method demonstrates lower decoding error rates compared to WavMark, AudioSeal, and SilentCipher, showing enhanced watermark durability under attack.

B Localization Refinement Module

To enhance the decoder's ability to accurately localize watermarked regions and reduce false detections, we introduce two replacement strategies during training:

Mismatched Watermark Replacement: For each watermarked audio sample, we randomly replace 85% of its embedded watermark segments with segments carrying different (non-target) watermark messages. This helps prevent the decoder from memorizing fixed patterns and promotes generalization across diverse watermark structures.

Random Segment Perturbation: We divide the audio into K segments and randomly select starting points for content replacement. Each selected segment (of length $2T/K$) is then altered with one of the following: 40% probability of clean (unwatermarked) waveform insertion, 20% probability of substitution with unrelated audio, and 20% probability of silence padding. The remaining 20% is left unchanged. These manipulations simulate realistic confusion patterns that the decoder may encounter in practice.

By combining these techniques and optimizing using the decoding loss $\mathcal{L}_{\text{msg}}$, the decoder is explicitly trained to focus on truly watermarked regions and reject irrelevant or misleading segments, significantly improving localization reliability during inference.

C Robustness Enhancement Module

To improve the watermark's resilience against signal processing attacks, we introduce a robustness enhancement module composed of 11 commonly used audio transformations. These operations are applied stochastically during training, with their parameters drawn from calibrated ranges. This helps the decoder learn to preserve watermark fidelity under real-world distortions.

Below we describe each transformation and its parameterization:

1. **Bandpass Filter** Removes both low- and high-frequency components while preserving

a specific mid-frequency range. *Parameters:* center frequency = 2750 Hz, quality factor $Q = 0.707$

2. **Highpass Filter** Attenuates frequencies below the cutoff, simulating microphone or channel filtering. *Parameters:* cutoff frequency = 1500 Hz

3. **Lowpass Filter** Attenuates frequencies above the cutoff, emulating bandwidth-limited scenarios. *Parameters:* cutoff frequency = 500 Hz

4. **Speed Adjustment** Alters playback speed by resampling, affecting both pitch and timing. *Parameters:* speed factor $\in$ [0.8, 1.2]

5. **Resampling** Converts to an intermediate sampling rate and back, introducing temporal interpolation artifacts. *Parameters:* resampled to 32kHz and then resampled back to the original frequency

6. **Boost** Multiplies the audio amplitude to simulate volume spikes or clipping. *Parameters:* boost factor = 10

7. **Duck** Reduces signal amplitude to mimic audio underpowering or suppression. *Parameters:* duck factor = 0.1

8. **Echo** Adds delayed and scaled versions of the signal to simulate reverberation. *Parameters:* delay time $\in$ [0.1, 0.5] seconds, echo volume $\in$ [0.1, 0.5]

9. **Pink Noise** Adds pknk noise to simulate natural ambient environments. *Parameters:* target SNR = 20 dB

10. **White Noise** Adds flat- Gaussian noise, resembling synthetic interference. *Parameters:* target SNR = 20 dB

11. **Smoothing** Applies a moving-average filter to blur waveform details. *Parameters:* window size $\in$ [2, 10] samples

Each transformation is sampled independently and applied with a certain probability during training. The combination and diversity of these perturbations guide the decoder to learn robust watermark recovery even under aggressive post-processing.

D TPR and FPR

First, we compared our method with the current state-of-the-art models (WavMark and AudioSeal) on several audio generation tasks, using True Positive Rate (TPR) and False Positive Rate (FPR) as evaluation metrics. TPR represents the proportion of watermarked audio correctly identified by the model, and its formula is:

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

where TP refers to true positives (samples correctly identified as watermarked) and FN refers to false negatives (samples with watermarks not detected). FPR represents the proportion of non-watermarked audio that is incorrectly classified as watermarked, and its formula is:

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

where FP refers to false positives (non-watermarked samples misclassified as watermarked) and TN refers to true negatives (samples correctly identified as non-watermarked). For watermarking models, our optimization goal is to maximize TPR while minimizing FPR.

E Detector Architecture

Inspired by the design of the AudioSeal watermark detector (San Roman et al., 2024), we implement a lightweight yet effective watermark detection model tailored for generative audio. Our detector operates directly on the raw audio waveform and outputs both a detection confidence and an optional binary message.

The architecture consists of two main components: an audio encoder and a classification head. The encoder, denoted as `self.encoder`, follows the same architectural design as the EnCodec encoder (Défossez et al., 2022), consisting of a series of downsampling convolutional blocks interleaved with residual connections. Specifically, the encoder comprises N convolutional layers with progressively increasing channel dimensions and strides to reduce temporal resolution, while preserving essential information for watermark detection. To restore alignment with the input resolution, a transposed convolution layer is applied after encoding.

Following the encoder, we apply a 1×1 convolution to produce a multi-head output. The first two

channels represent the confidence scores (via softmax) for the presence or absence of a watermark. The remaining n channels represent the per-bit logits of the embedded binary watermark message, which are decoded via a temporal average followed by a sigmoid activation. This design allows the detector to perform both binary watermark detection and payload recovery in a unified forward pass.

F Subjective Evaluation Protocol and Human Study Information

F.1 MUSHRA Test Setup

To evaluate the perceptual audio quality of watermarked audio, we conducted a subjective study using the standardized **MUSHRA protocol** (Multiple Stimuli with Hidden Reference and Anchor), following ITU-T Recommendation BS.1534-1. This method is widely used in audio quality benchmarking and provides robust human preference data across fine-grained quality levels.

Each test session included:

- One fixed reference audio clip (original unwatermarked audio),
- Three watermarked outputs (GenMark, AudioSeal, WavMark),
- Two lossy anchors: Anchor70 (band-limited at 7 kHz), Anchor35 (band-limited at 3.5 kHz),
- One hidden reference (identical to the original, included to assess rating consistency).

Participants evaluated the samples using an interactive web-based MUSHRA interface that supports waveform visualization, looping playback, and blind randomized ordering of stimuli. The interface was customized to guide the listener through the evaluation, showing condition names only during the training phase, and hiding them during formal scoring.

We recruited **20 expert listeners** with backgrounds in audio engineering or speech synthesis. All participants voluntarily agreed to take part in the study and were informed that their responses would be used for academic research purposes only. No personally identifying information (PII) was collected. As the evaluation involved non-sensitive, low-risk listening tasks, no formal IRB approval was required.

Each participant rated **20 audio groups**, each corresponding to a different prompt. Ratings were provided on a 0–100 scale via slider interfaces, with the ability to replay any sample as needed. Anchor and reference scores were used to validate listener consistency, and all results were aggregated by condition across listeners. For quantitative analysis and comparisons, please refer to Section 5.4 of the main paper.

The testing interface was implemented as a browser-based system supporting:

- Interactive MUSHRA scoring with waveform display and audio looping,
- Randomized presentation of audio conditions per trial,
- Automated anchor generation using standard low-pass filters.

F.2 Instructions Provided to Participants

Participants received the following instructions (translated and paraphrased from the interface):

> **Welcome to the Audio Quality Evaluation Test**
>
> This test assesses your subjective perception of audio quality.
>
> **Testing Process**:
>
> - Left panel: Reference audio (always visible)
> - Right panel: Six randomized test samples (three algorithmic outputs, two lossy anchors, one hidden reference)
>
> **Scoring Guide**:
>
> - 0–35: Severe degradation
> - 45–60: Moderate degradation
> - 61–80: Mild degradation
> - 80–100: Nearly indistinguishable from reference
>
> Please ensure a quiet environment and use high-quality headphones. Focus on high-frequency regions (e.g., fricatives like /s/, /z/) to detect perceptual differences.