Anchor Watermark: Robust Attribution for Diffusion-based Text-to-Audio Model

Xianjin Rong, Donghui Hu *

School of Computer Science and Information Engineering, Hefei University of Technology
rongxianjin@mail.hfut.edu.cn, hudh@hfut.edu.cn

Abstract

With the increasing commercialization of the latent diffusion-based text-to-audio generation, model attribution has become a critical challenge. Embedding watermarks in generated audio is an effective way to distinguish synthetic from natural audio. However, existing watermarking methods often suffer from limited robustness or require additional training, limiting their scalability in practical applications. In this paper, we propose an anchor-based inversion optimization framework. The method embeds a watermark into the model's initial latent vector, designated as a pivotal anchor, and extracts the watermark through inversion. To mitigate error accumulation and enhance robustness during inversion, we leverage the temporal consistency and distributional similarity of diffusion models, formulating watermark extraction as a time-series optimization problem. Specifically, given a suspicious audio sample and a candidate model with a predefined anchor, we first perform unguided denoising diffusion on the anchor to generate an intermediate latent trajectory as the anchor sequence. Then, we optimize the inversion process to align the inverted trajectory with the anchor sequence, thereby reducing accumulated errors. During optimization, we adopt Soft Dynamic Time Warping as the loss function. Its flexible temporal alignment capability ensures that correct attribution is achieved only when the anchor matches the target audio. Experimental results show that our method enables training-free attribution while preserving audio quality and achieving strong robustness.

Code — https://github.com/DDAN-LAB/AnchorWM.

Introduction

Latent diffusion models (LDMs) have recently gained substantial attention and have shown strong capabilities in text-to-audio (TTA) generation (Liu et al. 2023; Ghosal et al. 2023; Liu et al. 2024b; Majumder et al. 2024). While making personalized audio creation accessible, they also raise serious ethical and legal concerns. For instance, voice imitation can enable impersonation-based fraud, and synthetic media may facilitate misinformation. Thus, tracing the source of generated audio has become critically important.

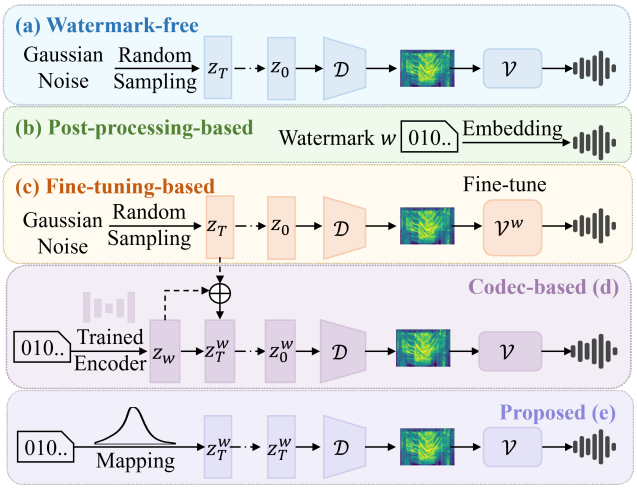


Figure 1: Existing methods fall into three categories: post-processing, fine-tuning, and codec-based approaches. In contrast, our method requires no additional training and leaves the model's generative performance unaffected.

Embedding watermarks into generated audio has emerged as the prevailing approach for attribution. Existing methods fall into three categories (Fig. 1): (1) post-processing approaches (Chen et al. 2023; Liu et al. 2024a) embed watermarks after generation but degrade audio quality; (2) vocoder fine-tuning methods (Cheng et al. 2024), which integrate watermarks into a fine-tuned HiFi-GAN (Kong and Kim 2020) but lack scalability across vocoder architectures; and (3) codec-based methods (Liu et al. 2024c) embed watermarks during generation using trained encoder-decoder models, which require additional training. All these approaches exhibit limited robustness, remaining vulnerable to perturbations such as resampling and pitch shifting.

Inspired by audio editing (Varshavsky-Hassid 2024) and image watermarking (Wen et al. 2023; Yang et al. 2024), which exploit diffusion reversibility by progressively adding noise to recover initial latent, we attempt to embed watermarks into the initial latent of TTA models without modifying the generative model and extract them by inverting the generated audio. However, our experiments show that direct inversion lacks robustness.

*Corresponding Author.

The primary challenge stems from the lossy conversion between mel-spectrograms and waveforms. This inherent part of the audio LDMs' architecture distorts the latent representation prior to inversion. Additionally, mismatches between inversion and the forward denoising process introduce further errors, which accumulate and eventually prevent watermark recovery. Notably, the intermediate latent vectors generated during denoising diffusion and inversion form two corresponding latent-space time series. We observe that when initialized with the same latent vector, the latent trajectories of classifier-free guided (CFG) diffusion (Ho and Salimans 2022) and unconditional diffusion remain closely aligned across time steps. Building on this insight, we propose Anchor Watermark, an anchor-based inversion optimization framework for robust attribution.

During embedding, we employ a distribution-preserving technique (Chen et al. 2022; Yang et al. 2024) to map the watermark into the model's initial latent space, forming a pivotal anchor, which makes the generated audio inherently carries the watermark (Fig. 1(d)). During extraction, we leverage temporal consistency in LDMs and the similarity between unguided and CFG diffusion to reformulate robustness as a time-series optimization problem. We first generate an anchor sequence by diffusing from the pivotal anchor, and treat the inverted latent sequence as the target for refinement. Soft Dynamic Time Warping (Soft-DTW) loss (Cuturi and Blondel 2017; Krause, Weiß, and Müller 2023) is then used to enforce temporal alignment between the two sequences. Attribution is deemed successful when the bit accuracy between the optimized latent watermark and the anchor surpasses a threshold. Experiments show that Anchor Watermark achieves strong attribution accuracy and robustness. Our contributions are summarized as follows:

1. We propose a novel attribution framework for latent TTA models that embeds watermarks into the initial latent and extracts them via audio inversion, without modifying the generation process or requiring additional training.
2. To ensure robustness, we optimize the inversion process to reduce error accumulation. Leveraging the trajectory similarity between unguided diffusion and CFG diffusion, we designate the initial watermarked latent as a pivotal anchor and employ Soft-DTW to enforce precise temporal alignment, ensuring that attribution succeeds only when the correct anchor is provided.
3. Experimental results on three datasets and four TTA models demonstrate superior effectiveness and robustness compared with state-of-the-art audio watermarking.

Overview

The overall framework is illustrated in Fig. 2. In our setting, each generated audio is embedded with a unique watermark w, where both the model parameters and the watermarking mechanism are accessible only to the model owner. Specifically, we employ a distribution-preserving technique (Yang et al. 2024; Chen et al. 2022) to map the watermark into the initial latent space, which serves as the pivotal anchor (prior evidence). For attribution, we first perform unguided diffusion on the pivotal anchor to produce an intermediate latent sequence that serves as optimization anchors. Next, the audio is inverted, and its latent vectors are optimized using these anchors. Finally, the watermark (posterior result) is recovered from the optimized inversion latents. If $Acc(w, w') > \tau$, the prior evidence and posterior result are considered matched, indicating successful attribution.

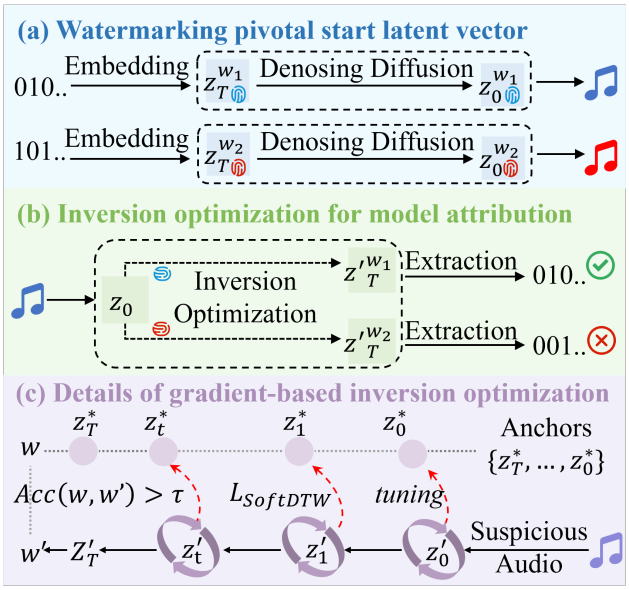


Figure 2: Overview of the proposed method.

Watermarking Pivotal Anchor

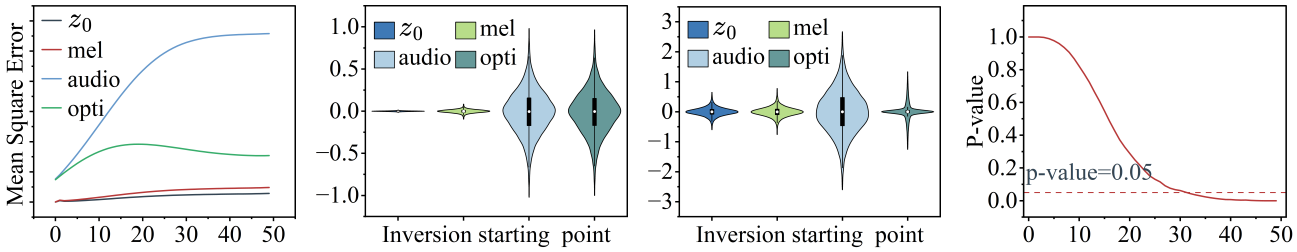
Watermark Embedding

Watermark embedding is achieved by mapping binary watermark bits into Gaussian noise. Specifically, we first reshape the watermark bits $w \in \{0,1\}$ to match the dimensionality of the random latent variable z_T. These bits are then encrypted using the ChaCha20 algorithm (Bernstein et al. 2008), producing a pseudo-random watermark w_r. Conditioned on w_r, we perform distribution-preserving sampling (Yang et al. 2024; Chen et al. 2022). Let $f(x)$ denote the probability density function of the Gaussian distribution $\mathcal{N}(0, I)$, and let ppf represent its percent-point function. We partition the density $f(x)$ into two regions of equal cumulative probability, where the bit value $y \in \{0,1\}$ follows a discrete uniform distribution, i.e., $p(y) = \frac{1}{2}$. Under the condition $y = i$, the watermarked latent variable z_T^w is sampled from the following distribution:

$$p(z_T^w | y = i) = \begin{cases} 2 \cdot f(z_T^w) & ppf(\frac{i}{2}) \le z_T^w \le ppf(\frac{i+1}{2}) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Denote the cumulative distribution function of $f(x)$ as cdf, the cumulative distribution function of Eq. 1 is:

$$F(z_T^w | y_i) = \begin{cases} 0 & z_T^w < ppf(\frac{i}{2}) \\ 2 \cdot cdf(z_T^s) - i & ppf(\frac{i}{2}) \le z_T^s \le ppf(\frac{i+1}{2}) \\ 1 & z_t^w > ppf(\frac{i+1}{2}). \end{cases} \tag{2}$$



(a) Inversion Error (b) Inversion timestep 1 (c) Inversion timestep 50 (d) K-S Test

Figure 3: (a) shows the MSE between latent vectors during inversion and denoising, while (b) and (c) report the corresponding variance changes. The results indicate errors grow progressively with diffusion steps, and the accumulation is more pronounced when additional conversions are involved. The optimization curves confirm that refining the inversion process effectively reduces errors. (d) presents a K-S test comparing latent distributions from classifier-free guided diffusion ($\lambda = 7.5$, $c \neq \emptyset$) and unguided diffusion($\lambda = 1$, $c = \emptyset$); the two distributions remain statistically consistent at most steps under a 0.05 threshold.

Let $u = F(z_T^w|y_i) \sim \mathcal{U}(0,1)$, the process of sampling the watermarked latent representation z_T^w driven by the randomized watermark, expressed as: $z_T^w = ppf\left(\frac{u+i}{2}\right)$. Please note that, in our setup, the encryption key (Bernstein et al. 2008) and model parameters are exclusively held by the model owner. Users can access the audio generation service only via a public API.

Audio Generation

Starting with watermarked z_T (designated as the pivotal anchor), TTA model iteratively performs deterministic denoising diffusion (Song, Meng, and Ermon 2021; Song et al. 2021) T steps: $z_0 = \mathcal{T}_\theta(z_T)$ with a classifier-free guidance (Ho and Salimans 2022). Then, the Mel-spectrogram is reconstructed using decoder: $m = \mathcal{D}(z_0)$. Finally, the vocoder converts the Mel-spectrogram back into an audio signal: $x = \mathcal{V}(m)$. The process above is identical to that of a non-watermarked model. For clarity of presentation, we refer to the noising process as inversion ($z_0 \rightarrow z_T$) and the denoising (generation inference) process as diffusion ($z_T \rightarrow z_0$).

Watermark Extraction

To extract the watermark, we first convert the audio x into a Mel-spectrogram $m = \mathcal{M}(x)$ and encode it into the latent space as $z_0^{'} = \mathcal{E}(m)$. Ideally, DDIM inversion (Song, Meng, and Ermon 2021) yields an approximate latent vector $z_T^{'} = \mathcal{T}_\theta^{-1}(z_0^{'}) \approx z_T^w$. The watermark is then extracted from the inverted latent $z_T^{'}$ via the inverse mapping function: $w' = \lfloor 2 \cdot cdf(z_T^{'}) \rfloor$. However, our experiments reveal that direct inversion results in poor watermark robustness.

Pivotal Anchor-based Attribution

Robustness Analysis

The lossy waveform-Mel conversion. Recall that audio generation and inversion require converting between waveforms and Mel-spectrograms, an inherently lossy process. The transformation $x \rightarrow m$ applies STFT followed by a Mel filter bank, which preserves low-frequency detail but suppresses high-frequency components, discarding spectral information. The reverse conversion $m \rightarrow x$ is even more lossy due to missing phase and further high-frequency attenuation. Consequently, the inverted latent z_0' deviates markedly from the original watermarked latent z_0^w, causing inevitable watermark degradation during state conversion.

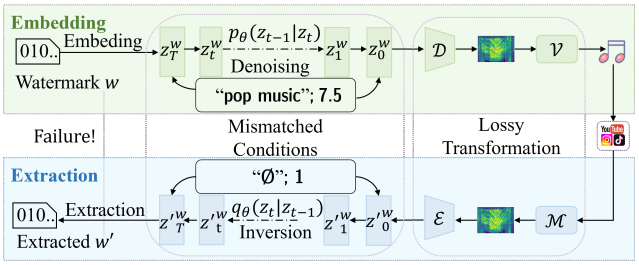


Figure 4: The explanation of the failure extraction through direct inversion.

Inversion error accumulation. During DDIM inversion, the diffusion process is reversed step by step, accumulating small errors at each stage. The denoising process uses classifier-free guidance (CFG) (Ho and Salimans 2022) with $\lambda > 1$ and a prompt $c \neq \emptyset$:

$$\overrightarrow{\epsilon_\theta}(z_t, t, c) = \lambda \cdot \epsilon_\theta(z_t, t, c) + (1 - \lambda) \cdot \epsilon_\theta(z_t, t, \emptyset). \quad (3)$$

However, the inversion stage lacks such conditions and estimates noise with $\lambda = 1$ and $c = \emptyset$: $\overleftarrow{\epsilon_\theta}(z_t, t) = \epsilon_\theta(z_t, t, \emptyset)$. The condition mismatch between the forward and reverse processes, combined with the degradation of the initial vector z_0' caused by lossy conversion, leads to cumulative errors during inversion. Fig. 3 reports the quantitative analysis of robustness, while Fig. 4 presents an intuitive depiction of the process of watermark audio generation and inversion-based extraction.

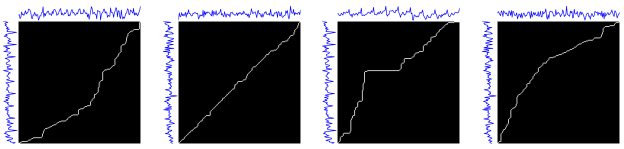
Pivot-based Attribution

Pivotal diffusion provides anchor sequence. During attribution, we first perform unconditional denoising diffusion on the pivot z_T^w using parameters $\lambda = 1$ and $c = \emptyset$, which

is consistent with the inversion configuration. This process yields a deterministic latent trajectory $\mathcal{Z}^* = \{z_T^*, \ldots, z_0^*\}$, where $z_T^w = z_T^{'}$ is fixed. The intermediate latent variables z_i along this trajectory closely resemble the distributions z_i^w obtained in the original generation process with $w > 1$ and $c \neq \emptyset$. These variables serve as reliable anchors for optimization, and effectively reduce inversion errors (see Fig. 3).

Leveraging distribution similarity. As previously mentioned, when the audio x is generated through a denoising diffusion process based on the pivotal anchor latent z_T^w, its inversion trajectory $\mathcal{Z}^{'} = \{z_0^{'}, \ldots, z_T^{'}\}$ is expected to exhibit strong similarity to the anchor trajectory $\mathcal{Z}^*$. Leveraging this similarity, we optimize the inverted latent vectors using the anchor sequence during the inversion process to reduce inversion errors and achieve accurate attribution. Specifically, at each time step $z_t^{'}$ and its corresponding anchor z_t^* :

$$\min ||\{z^*\}_{t=0}^T - \{z^{'}\}_{t=0}^T||. \tag{4}$$



(a) $p(z_1^*, z_1')$ (b) $p(z_{50}^*, z_{50}')$ (c) $p(z_1^*, \hat{z}_1)$ (d) $p(z_{50}^*, \hat{z}_{50})$

Figure 5: SoftDTW alignment paths. We optimize the watermarked latent trajectory $z^{'}$ and the non-watermarked latent trajectory $\hat{z}$ with respect to the anchor z^*. The results reveal distinct alignment paths between the two cases.

SoftDTW for temporal alignment. To balance watermark robustness with attribute accuracy, we adopt Soft Dynamic Time Warping (SoftDTW) (Cuturi and Blondel 2017) as the loss function. Unlike Mean Squared Error (MSE), which only measures pointwise differences, SoftDTW probabilistically evaluates all possible alignment paths, enabling flexible temporal alignment between sequences. This characteristic makes it particularly well-suited for optimizing latent diffusion sequences, as diffusion and inversion trajectories can exhibit temporal misalignments due to lossy transformations. As illustrated in Fig. 5, SoftDTW yields distinctly different alignment paths for watermarked versus non-watermarked audio. Mathematically, SoftDTW calculates the cumulative alignment cost between two sequences by recursively updating a cost matrix R:

$$R_{i,j} = D_{i,j} + \text{softmin}_\gamma\big(R_{i-1,j}, R_{i,j-1}, R_{i-1,j-1}\big), \tag{5}$$

where $D_{i,j} = \|z_i - z_j'\|^2$ denotes the pairwise distance matrix, and softmin_γ is defined as:

$$\text{softmin}_\gamma(\mathbf{a}) = -\gamma \cdot \log \sum_k \exp(-a_k/\gamma), \tag{6}$$

where $\gamma > 0$ is a smoothing parameter. The final SoftDTW loss is defined as: $\mathcal{L}_{\text{SoftDTW}}(z^*, z') = R_{i,j}$.

Gradient-based inversion optimization. At each step $t \in \{0, 1, \ldots, T\}$, we compute the SoftDTW loss between z_t^* and z_t', and update z_t' via gradient descent (Fig. 2(c)). As optimization progresses, the optimized trajectory $\mathcal{Z}' = \{z_0', \ldots, z_T'\}$ gradually approaches the anchors $\mathcal{Z}^* = \{z_0^*, \ldots, z_T^*\}$. This improves the final inverted latent z_T' for watermark extraction. The SoftDTW loss at each iteration is computed as:

$$\mathcal{L}_t(z_t^*, z_t') = \text{SoftDTW}\big(\{z_t^*\}_{t=0}^T, \{z_t'\}_{t=0}^T\big). \tag{7}$$

The gradient $\frac{\partial \mathcal{L}}{\partial D_{i,j}}$ can be computed via backward dynamic programming using the normalized alignment probabilities. By the chain rule, the gradient for each latent element is computed as:

$$\frac{\partial \mathcal{L}_t}{\partial z_i} = \sum_{j=1}^m \frac{\partial \mathcal{L}_t}{\partial D_{i,j}} \cdot \frac{\partial D_{i,j}}{\partial z_i} = 2 \sum_{j=1}^m P_{i,j}(z_i - z_j'), \tag{8}$$

where $P_{i,j}$ is the alignment probability derived from the backward pass. We then perform gradient descent with the learning rate η to update each z_t' :

$$z_t' \leftarrow z_t' - \eta \cdot \frac{\partial \mathcal{L}_t}{\partial z_t'}. \tag{9}$$

The updated latent z_t' is then used as the output for subsequent inversion iteration. By leveraging stable and reliable anchors, the model progressively computes a more accurate inverted latent z_{t+1}'. This iterative refinement effectively mitigates cumulative inversion errors and yields more robust watermark properties compared to direct inversion.

Experimental Results and Analysis

Experimental Setup

TTA models & Datasets. We conduct experiments on four state-of-the-art open-source T2A diffusion models: AudioLDM (Liu et al. 2023), AudioLDM2 (Liu et al. 2024b), Tango (Ghosal et al. 2023), and Tango2 (Majumder et al. 2024). We perform watermarked audio generation on three benchmark datasets: AudioCaps (Kim and Kim 2019), WavCaps (Mei, Meng, and and 2024), and MusicCaps (Agostinelli et al. 2023).

Baseline methods. We perform a comprehensive comparison of the proposed method with existing approaches, including post-processing watermarking methods (FSVC (Zhao et al. 2021), NormSpace (Saadi, Merrad, and Benziane 2019), Patch (Natgunanathan et al. 2017), AudioSeal (Roman et al. 2024), WavMark (Chen et al. 2023), and Timbre (Liu et al. 2024a)), the fine-tuning-based watermarking method HiFi-GANw (Cheng et al. 2024), and the codec-based watermarking method GROOT (Liu et al. 2024c).

Distortions method. To assess watermark robustness, we subjected the watermarked audio to ten types distortions: 32 kHz resampling (RS), 100 Hz high-pass filtering (HP), 1000 Hz low-pass filtering (LP), 10% amplitude scaling (AS), 9 kbps MP3 compression (MC), 8 bps recount (RC), median filtering with a 15-sample window (MF), 5 dB Gaussian noise (GN), 60% cropping (CB), and a semitone(0.5) pitch shift (PS).

Method	Distortions												FAD	IS
	Clean	CP	RS	GN	AS	MC	RC	MF	LF	HF	PS	Avg	*(t-value↓)*	*(t-value↓)*
AudioLDM	-	-	-	-	-	-	-	-	-	-	-	-	2.32	7.57
Fsvc	**1.00**	0.05	**1.00**	0.26	**1.00**	0.75	0.28	0.80	0.97	**1.00**	0.21	0.63	2.83	7.47
	1.00	0.69	**1.00**	0.71	0.99	0.82	0.67	0.81	0.97	**1.00**	0.69	0.84	*9.59*	*0.34*
Patch	**1.00**	0.04	**1.00**	0.00	0.91	0.00	0.00	0.00	0.00	**1.00**	0.00	0.30	3.20	7.46
	0.85	0.63	0.84	0.58	0.81	0.54	0.59	0.53	0.53	0.84	0.51	0.64	*8.91*	*1.49*
Norm	**1.00**	0.00	**1.00**	0.90	**1.00**	0.96	0.00	0.96	0.97	0.87	0.00	0.67	2.86	7.34
	1.00	0.69	**1.00**	0.89	**1.00**	0.92	0.49	0.92	0.94	0.90	0.51	0.83	*8.74*	*0.39*
Timbre	**1.00**	**1.00**	**1.00**	0.00	**1.00**	0.00	0.00	0.57	0.25	**1.00**	0.00	0.48	3.10	6.89
	1.00	**1.00**	**1.00**	0.59	**1.00**	0.63	0.51	0.75	0.73	**1.00**	0.54	0.78	*8.74*	*2.51*
WavMark	**1.00**	0.59	**1.00**	0.00	**1.00**	0.02	0.00	0.25	0.01	**1.00**	0.00	0.39	5.83	5.35
	1.00	0.75	**1.00**	0.50	0.99	0.51	0.50	0.64	0.49	**1.00**	0.47	0.68	*15.78*	*3.34*
AudioSeal	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	0.00	0.20	3.10	7.47
	0.94	0.56	0.69	0.38	0.56	0.50	0.44	0.56	0.88	**1.00**	0.63	0.62	*10.51*	*0.48*
HiFI-GANw	**1.00**	**1.00**	**1.00**	0.00	**1.00**	0.00	0.00	0.57	0.24	**1.00**	0.00	0.48	3.24	6.91
	1.00	**1.00**	**1.00**	0.56	**1.00**	0.63	0.54	0.75	0.73	**1.00**	0.54	0.77	*8.77*	*2.77*
GROOT	**1.00**	0.84	0.99	0.91	**1.00**	**1.00**	0.07	0.97	**1.00**	0.99	0.46	0.82	2.57	7.23
	1.00	0.91	0.99	0.89	**1.00**	**1.00**	0.53	0.99	0.99	**1.00**	0.72	0.90	*18.32*	*1.36*
Ours	**1.00**	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	2.31	7.59
	1.00	**1.00**	**1.00**	**0.95**	**1.00**	0.99	**0.98**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	***1.68***	***0.30***

Table 1: Comparison results. In each row, the top reports the TPR and the bottom shows the Bit Acc. Model bias is assessed via a t-test on mean FAD and IS against watermark-free audios, with t-values given in *italics*. Best results are in **bold**.

Evaluation metrics. We compute the true positive rate (TPR) at a fixed false positive rate (FPR) and evaluate the bit accuracy (Bit Acc) of the watermark. Furthermore, we exploit Fréchet Audio Distance (FAD) and inception score (IS) to assess the quality of the audio, t-value is employed to evaluate the performance bias introduced by watermarking.

Diffusion and inversion setting. During generation, the DDIM (Song, Meng, and Ermon 2021) sampler performs 50 steps to produce 8s audio that embeds a 256-bit watermark. For inversion, we use DDIM inversion with an empty prompt ($c = \emptyset$) and a guidance scale of 1. Optimization is performed using the Adam optimizer (lr = 0.01) with three iterations per timestep. All values are rounded to the nearest unit. The experiments are implemented in PyTorch on an RTX 4090 GPU.

Attribution Performance

We assume there are N anchor watermarks $w_i \in 0, 1$ $(i = 1, \ldots, N)$, the suspicious audio is optimized with each anchor N_i, and the bit-match $\text{Acc}(w_i, w')$ with the extracted watermark w' is computed. If all tests fall below threshold τ, the audio is considered non-watermarked; otherwise, it is model-generated, and $\arg\max_i \text{Acc}(w_i, w')$ is taken as the attribution. For a given τ, the false positive rate (FPR) is defined as: $\text{FPR}(\tau, N) = 1 - (1 - \text{FPR}(\tau))^N \approx N \cdot \text{FPR}(\tau)$ (Fernandez and and 2023). In our main experiments, we control FPR at 10^{-15} and set N to 10^{10}, calculates a threshold $\tau = 0.8164$. As shown in Tab. 2, our method achieves consistently performance across all four TTA models and three datasets. The TPR reaches 100% on clean audio, and only minor drops under attacks, demonstrating excellent performance while maintaining strong robustness.

Comparison to Baselines

Robustness against distortions. We employ the AudioLDM model with prompts from the AudioCaps dataset to generate 1, 000 watermarked audio samples for each method. As Tab. 1 shows, most methods perform well on clean audio but degrade sharply under cropping (CP), Gaussian noise (GN), or re-encoding (RC). Patch, Norm, Timbre, WavMark, and AudioSeal fail in most attacks, while our method maintains high robustness with average TPR 98% and Bit Acc 99%. Although GROOT performs better than other baselines, it still lags with TPR 16% and Bit Acc 19%, confirming the superior robustness of our approach.

Performance bias. To assess the model's performance degradation caused by watermark embedding, we conducted a t-test. The hypotheses are defined as $H_0 : \mu_s = \mu_0$ and $H_1 : \mu_s \neq \mu_0$, where μ_s and μ_0 represent the average FAD or IS computed on watermarked and non-watermarked audios (10 groups of 1,000 samples each), respectively. A smaller t value indicates a higher probability of H_0 being true, implying no significant performance degradation. Conversely, if the t value exceeds a predefined threshold, H_0 is rejected, i.e., model performance is affected. As shown in Tab. 1, our method achieves the lowest t value, suggesting

TTA Model	AudioCaps				MusicCaps				WaveCaps			
	TPR		Bit Acc		TPR		Bit Acc		TPR		Bit Acc	
AudioLDM	1.000	0.983	1.000	0.992	1.000	0.991	1.000	0.994	1.000	0.986	1.000	0.990
AudioLDM 2	1.000	1.000	1.000	0.995	1.000	0.980	1.000	0.988	1.000	0.980	1.000	0.988
Tango	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
Tango 2	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999

Table 2: Performance of the method. The left value in each column corresponds to the result on clean watermarked audio, while the right corresponds to the result on attacked watermarked audio. The results demonstrate that the proposed watermarking scheme achieves high accuracy, strong robustness, and good generalization capability across different TTA models and datasets.

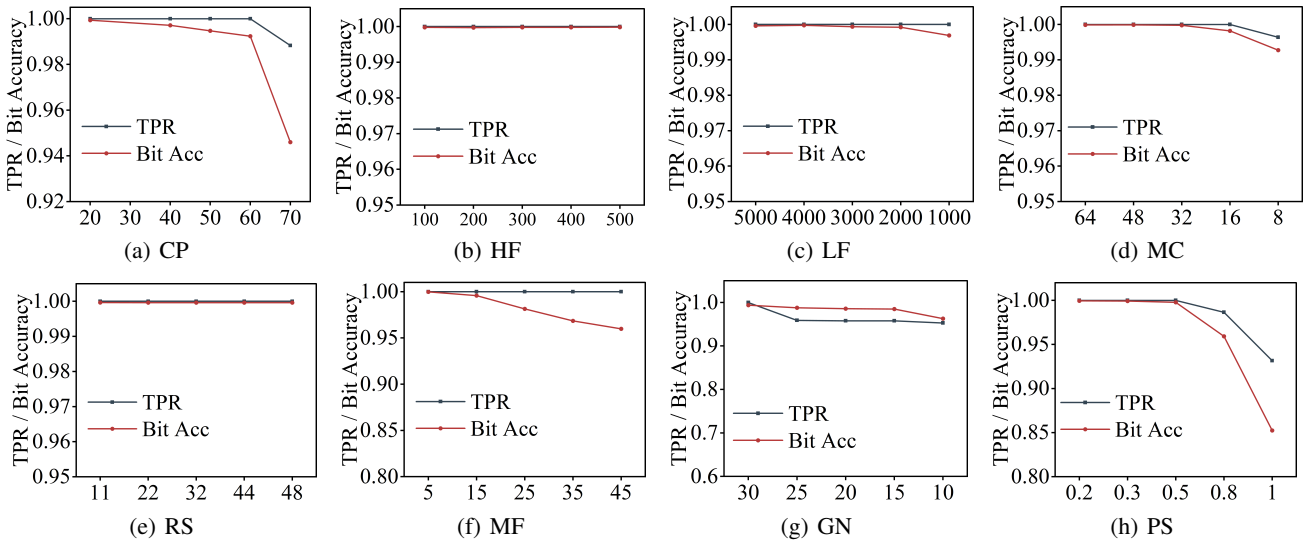


Figure 6: Robustness against different distortion intensities.

that the audio quality is effectively preserved.

Optimization Settings

Effectiveness of optimization. To assess the effectiveness of the proposed inversion optimization scheme, we compared it against direct DDIM inversion (Song, Meng, and Ermon 2021). Figure 7(a) shows the average performance of the two inversion methods under the perturbations described in Distortions Method section. Compared with direct inversion, our optimized inversion improves the TPR and bit accuracy by 10% and 7% on AudioLDM (Fig. 7(a), left), and by 6% and 5% on Tango (Fig. 7(a), right), demonstrating the proposed method significantly enhances robustness.

Loss function. As shown in tab. 3 (left), SoftDTW achieves higher and more stable accuracy on watermarked audio. In contrast, Cross Entropy shows poor robustness, while MSE and KLD perform poorly on both generated and natural audio. This is because SoftDTW is mathematically better suited for time-series optimization: it supports flexible temporal alignment, and provides a smooth and stable optimization process, achieving superior performance.

Learning rate. From Tab. 3 (middle), we conclude that low (0.005) and medium (0.05) learning rates yield high inversion accuracy, especially on watermarked audio. In contrast, a high learning rate (0.1) reduces accuracy on generated and natural audio, indicating that excessively large optimization steps lead to instability and unreliable results.

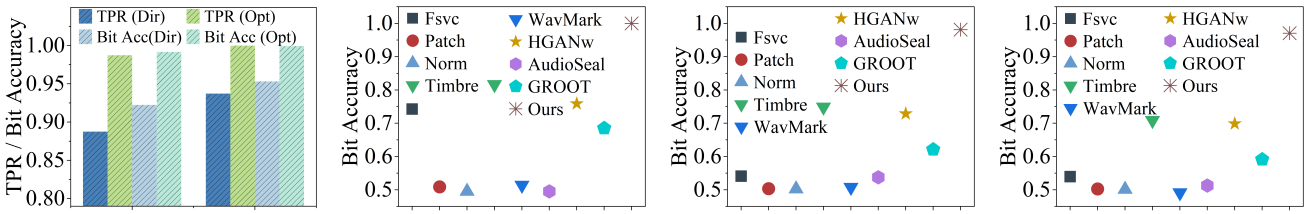
Iteration steps. Tab. 3 (right) indicates that increasing inner steps from 2 to 8 yields little accuracy improvement and slightly worse performance on natural audio, while detection time rises, from approximately 4s to 17s per audio. More iteration steps increase the time cost without clear benefit, indicating our method can balance accuracy and efficiency.

Adaptive Attacks

We evaluated three attack scenarios: Mel-spectrogram reconstruction ($audio \rightarrow Mel \rightarrow audio$), VAE compression ($audio \rightarrow Mel \rightarrow z_0 \rightarrow Mel \rightarrow audio$), and regeneration attack ($audio \rightarrow Mel \rightarrow z_0 \rightarrow\rightarrow z_T \rightarrow z_0 \rightarrow Mel \rightarrow audio$). For Mel-spectrogram reconstruction, STFT was applied directly. To simulate the security risks posed by model leakage, we used the same AudioLDM model in our experiments (noting that in practical applications, model parameters are generally inaccessible). Accordingly, in VAE compression we adopted the parameters of AudioLDM, and in the regeneration attack we also employed this model by

Audio type	Loss function				Learing rate			Iteration step		
	SoftDTW	MSE	KLD	Cross	0.005	0.05	0.1	2	5	8
Clean	1.00/1.00	1.00/1.00	0.00/0.54	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
Distortion	0.98/0.99	1.00/0.98	0.00/0.52	0.84/0.92	0.97/0.98	1.00/0.99	1.00/0.99	0.97/0.98	1.00/0.99	1.00/0.99
Generated	1.00/0.57	0.00/0.96	1.00/0.51	1.00/0.48	1.00/0.55	0.84/0.71	0.83/0.73	1.00/0.56	1.00/0.64	0.93/0.69
Natural	1.00/0.61	0.00/0.97	1.00/0.49	1.00/0.50	1.00/0.56	0.83/0.76	0.81/0.77	1.00/0.57	1.00/0.68	0.81/0.74

Table 3: The performance (TPR/Bit Acc) under different optimization settings. Clean denotes undistorted watermarked audio. Distortion represents the average result across all attacked watermarked audio. Generated refers to watermark-free audio generated by MusicTango (Agostinelli et al. 2023). Natural indicates real audio from the AudioCaps dataset.



(a) Effectiveness (b) Reconstruct (c) Compress (d) Regenerate

Figure 7: The effectiveness of the proposed optimization method against common distortions (a) and adaptive attacks (b-d).

Settings	Guidance scale				Inversion step				Sampling method			
	4	10	14	18	10	25	75	100	LCM	PNDM	DEIS	DPMSolver
Clean	1.000	1.000	0.997	0.997	0.997	0.999	0.999	0.999	1.000	1.000	0.997	0.997
Distortion	0.995	0.984	0.984	0.982	0.986	0.990	0.991	0.985	0.996	0.997	0.997	0.990

Table 4: The bit accuracy with different sampling settings.

first inverting the audio to obtain its latent representation, followed by denoising with an empty prompt to regenerate the audio and extract the watermark. As illustrated in Fig. 7, our method consistently and reliably recovers the watermark even under these three erasure attacks. From a security standpoint, adversaries could potentially steal the model and use arbitrary prompts to generate harmful audio, emphasizing the critical importance of preventing model leakage.

Ablation Studies

Distortions intensities To further evaluate robustness, we conducted experiments under different perturbation intensities. As shown in the Fig. 6, for pitch shift, the performance drops significantly at higher intensities; however, it still maintains a bit accuracy of about 85% and a TPR of 93%. For the other distortions, the results remain above approximately 94% even under high-intensity conditions.

Guidance scale, inversion steps and sampling methods. To assess the generalization capability of the proposed method, we varied the guidance scale (from 4 to 18), the sampling steps (from 10 to 100), and the ODE samplers (LCM (Luo et al. 2023), PNDM (Liu et al. 2022), DEIS (Zhang and Chen 2023), and DPMSolver (Lu et al. 2022)) on AudioLDM. In each experiment, only one parameter was modified while the others were kept at their default settings. As shown in Tab. 4, the watermark was successfully and consistently extracted under all configurations, confirming the robustness and generalizability of the proposed method across diverse setups.

Conclusion and Future Work

We propose an anchor watermarking framework based on latent inversion optimization. The method embeds a watermark into the initial latent as a pivotal anchor and extracts it through the inversion process. To mitigate inversion errors, we leverage the temporal consistency and generative stability of diffusion models and formulate watermark extraction as a time-series optimization problem. Soft Dynamic Time Warping is introduced to enable flexible temporal alignment. Experimental results demonstrate that our method maintains robustness under a wide variety of attacks. Unlike existing approaches, our method requires no additional training and does not compromise generation quality. In real-world deployments, although the storage overhead of anchors is negligible, managing these anchors may still pose practical concerns. In future work, we plan to explore optimization techniques suitable for more open environments and further investigate the effectiveness of anchor-based sequence optimization for broader generative watermarking scenarios, including image and video generation.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J. H.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. H. 2023. MusicLM: Generating Music From Text. *CoRR*, abs/2301.11325.

Bernstein, D. J.; et al. 2008. ChaCha, a variant of Salsa20. In *Workshop record of SASC*, volume 8, 3–5. Citeseer.

Chen, G.; Wu, Y.; Liu, S.; Liu, T.; Du, X.; and Wei, F. 2023. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*.

Chen, K.; Zhou, H.; Zhao, H.; Chen, D.; Zhang, W.; and Yu, N. 2022. Distribution-Preserving Steganography Based on Text-to-Speech Generative Models. *IEEE Trans. Dependable Secur. Comput.*, 19(5): 3343–3356.

Cheng, X.; Wang, Y.; Liu, C.; Hu, D.; and Su, Z. 2024. HiFi-GANw: Watermarked Speech Synthesis via Fine-Tuning of HiFi-GAN. *IEEE Signal Processing Letters*, 31: 2440–2444.

Cuturi, M.; and Blondel, M. 2017. Soft-DTW: a Differentiable Loss Function for Time-Series. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 894–903. PMLR.

Fernandez, P.; and and, G. C. 2023. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 22409–22420. IEEE.

Ghosal, D.; Majumder, N.; Mehrish, A.; and Poria, S. 2023. Text-to-Audio Generation using Instruction Guided Latent Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 3590–3598. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Kim, C. D.; and Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132.

Kong, J.; and Kim, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In Larochelle, H.; and Ranzato, M., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Krause, M.; Weiß, C.; and Müller, M. 2023. Soft Dynamic Time Warping for Multi-Pitch Estimation and Beyond. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, 1–5. IEEE.

Liu, C.; Zhang, J.; Zhang, T.; Yang, X.; Zhang, W.; and Yu, N. 2024a. Detecting Voice Cloning Attacks via Timbre Watermarking. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26-March 1, 2024*. The Internet Society.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D. P.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 21450–21474. PMLR.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024b. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32: 2871–2883.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Liu, W.; Li, Y.; Lin, D.; Tian, H.; and Li, H. 2024c. GROOT: Generating Robust Watermark for Diffusion-Model-Based Audio Synthesis. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024-1 November 2024*, 3294–3302. ACM.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28-December 9, 2022*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. *CoRR*, abs/2310.04378.

Majumder, N.; Hung, C.; Ghosal, D.; Hsu, W.; Mihalcea, R.; and Poria, S. 2024. Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 564–572. ACM.

Mei, X.; Meng, C.; and and, H. L. 2024. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32: 3339–3354.

Natgunanathan, I.; Xiang, Y.; Hua, G.; Beliakov, G.; and Yearwood, J. 2017. Patchwork-Based Multilayer Audio Watermarking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11): 2176–2187.

Roman, R. S.; Fernandez, P.; Elsahar, H.; Défossez, A.; Furon, T.; and Tran, T. 2024. Proactive Detection of Voice Cloning with Localized Watermarking. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Saadi, S.; Merrad, A.; and Benziane, A. 2019. Novel secured scheme for blind audio/speech norm-space watermarking by Arnold algorithm. *Signal Processing*, 154: 74–86.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Varshavsky-Hassid, M. 2024. On the Semantic Latent Space of Diffusion-Based Text-to-Speech Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 246–255.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-rings watermarks: invisible fingerprints for diffusion images. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Yang, Z.; Zeng, K.; Chen, K.; and Fang, H. 2024. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 12162–12171. IEEE.

Zhang, Q.; and Chen, Y. 2023. Fast Sampling of Diffusion Models with Exponential Integrator. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhao, J.; Zong, T.; Xiang, Y.; Gao, L.; Zhou, W.; and Beliakov, G. 2021. Desynchronization Attacks Resilient Watermarking Method Based on Frequency Singular Value Coefficient Modification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2282–2295.