# Black-box Audio Watermark Removal via Diffusion Priors

**Lingfeng Yao** [1] **Xincong Zhong** [2] **Chenpei Huang** [1] **Xuandong Zhao** [3] **Hanqing Guo** [4]
**Aohan Li** [5] **Jiang Liu** [2] **Tomoaki Ohtsuki** [6] **Miao Pan** [1]

## Abstract

With the rise of AI-generated audio, watermarking has become widely used for detecting misuse and protecting intellectual property. However, adversaries may try to remove these watermarks, making it critical to evaluate how well watermarking schemes withstand removal attacks. Existing attacks are often impractical: they either noticeably degrade perceptual quality or require access to the watermarking scheme. We propose DiffErase, a black-box watermark removal attack that assumes no knowledge of the target watermarking scheme while maintaining perceptual quality. DiffErase perturbs watermarked audio to an intermediate diffusion noise level and regenerates it using a pretrained denoising model, effectively suppressing watermark signals. Theoretical analysis and extensive experiments demonstrate that inaudible audio watermarks are highly vulnerable: across multiple audio domains, DiffErase consistently removes watermarks while preserving perceptual quality. These findings highlight the need for future audio watermarking designs to consider diffusion-based threats. Code and demos are available at https://differase.github.io/DiffErase/.

## 1. Introduction

Audio generative models can now produce highly realistic audio that is indistinguishable to ordinary listeners. This technology reduces the cost of audio creation and enables many beneficial applications, but it also raises security risks related to misinformation, fraud, and public distrust. To mitigate these risks, proactive defenses such as audio watermarking have been proposed to support governance of AI-generated content (Roman et al., 2024; Singh et al., 2024). By embedding an imperceptible signal into generated audio, a watermark provides a provenance cue that can be verified later, enabling detection of AI-generated audio.

[1]University of Houston [2]Waseda University [3]UC Berkeley [4]University of Hawaii at Mānoa [5]The University of Electro-Communications [6]Keio University. Correspondence to: <>.

*Preprint. February 12, 2026.*

Nevertheless, motivated adversaries may try to remove or disable watermarks to evade such verification. This makes it necessary to evaluate the robustness of watermarking schemes against removal attacks, ensuring they can withstand real-world threats. Prior work has studied attacks against audio watermarking. Some works (Wen et al., 2025; O'Reilly et al., 2025) apply signal-level transformations (e.g., additive noise or pitch shifting) or physical-level operations (e.g., re-recording) to disable detection. However, effective removal under these operations often introduces audible distortion, limiting their practicality. Other works explore more targeted strategies. Adversarial attacks (Liu et al., 2024; Li et al., 2025) optimize perturbations added to the watermarked waveform to disable detection, but they typically require query access to detector output. Overwriting attacks (Yao et al., 2025; Liu et al., 2023b) embed an additional watermark to interfere with or replace the original, but they often depend on knowledge of the target watermarking system, such as the embedding architecture. These limitations raise a natural question: *Can an adversary remove audio watermarks in a black-box setting while preserving perceptual quality?*

In this work, we investigate diffusion-based regeneration as a black-box watermark removal strategy. The key insight is that watermarking adds imperceptible perturbations that shift samples away from the natural data distribution, which detectors are trained to recognize. Diffusion models, trained to denoise corrupted inputs back to the data manifold, may naturally suppress these structured perturbations while preserving perceptual quality. Importantly, this approach requires no knowledge of the watermarking scheme, making it fully black-box. This intuition is supported by recent work on regeneration attacks against image watermarks (Saberi et al., 2024; Zhao et al., 2024) and diffusion-based defenses against adversarial examples (Nie et al., 2022; Wu et al., 2023; Guo et al., 2024), where diffusion suppresses imperceptible perturbations. However, whether diffusion priors can enable black-box watermark removal for audio watermarking while preserving quality remains unexplored.

Extending diffusion-based regeneration from images to audio requires careful design choices, as audio admits multiple representations with different reconstruction constraints. A direct application of waveform-level diffusion (Kong et al.,

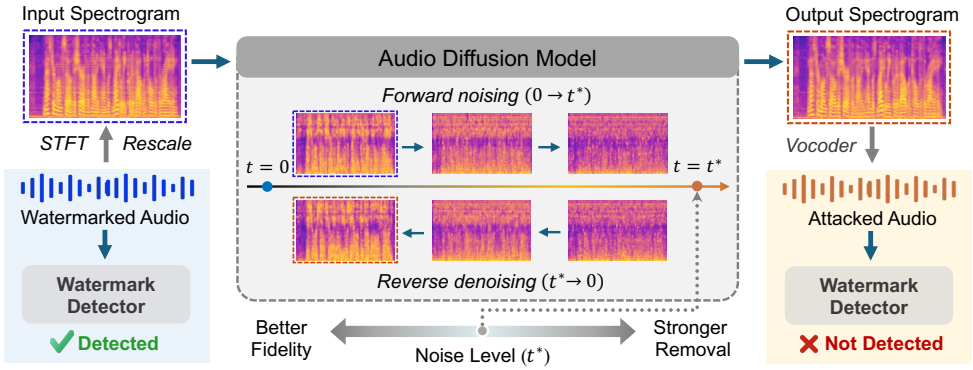



*Figure 1.* **Overview of DiffErase.** Watermarked audio is converted to a mel-spectrogram via STFT, perturbed to an intermediate noise level $t^*$ via forward noising, and then denoised back to $t = 0$. A vocoder reconstructs the attacked audio, which evades watermark detection. The noise level $t^*$ controls the trade-off between removal strength and perceptual fidelity.

2020b) leads to over-smoothed content and temporal drift, degrading perceptual similarity. Alternatively, one can diffuse a linear spectrogram and invert it using the original phase (Guo et al., 2024), but the regenerated magnitude may be inconsistent with the preserved phase, producing audible artifacts. *These issues make it challenging to obtain both strong watermark removal and high fidelity preservation.*

We propose DiffErase, a diffusion-based watermark removal attack operating in a strictly black-box setting. As illustrated in Figure 1, DiffErase converts watermarked audio to a mel-spectrogram using STFT, perturbs it to an intermediate noise level $t^*$ via the forward noising process, and applies a pretrained denoising model to regenerate the mel-spectrogram back to $t = 0$. A neural vocoder then reconstructs the waveform from the regenerated mel-spectrogram. The mel-spectrogram representation preserves salient perceptual structure such as energy contours and temporal envelopes, while enabling seamless integration with modern neural vocoders for high-quality reconstruction. We implement two complementary diffusion backbones: (i) mel-spectrogram diffusion operating directly in the mel-spectrogram domain, and (ii) latent diffusion works in a learned latent space from mel-spectrogram features.

To understand why DiffErase succeeds, we provide a manifold-based analysis that models watermark embedding as an off-manifold perturbation. We show that diffusion reverse dynamics contract the watermark residue along the denoising trajectory, with an exponential decay bound controlled by the noise level. We evaluate DiffErase across three audio domains (speech, music, and environmental sounds) against five state-of-the-art watermarking systems (AudioSeal (Roman et al., 2024), TimbreWM (Liu et al., 2023b), WavMark (Chen et al., 2023), Perth (Resemble AI, 2025), and SilentCipher (Singh et al., 2024)). Our results demonstrate that DiffErase consistently disables watermark detection while maintaining high perceptual quality, highlighting diffusion-based regeneration as a practical threat that must be addressed in future watermark designs.

Our contribution can be summarized as follows:

- We propose DiffErase, a black-box diffusion-based attack that removes neural audio watermarks without requiring knowledge of the watermarking schemes.
- We provide a theoretical analysis explaining why diffusion dynamics suppress inaudible watermarks, with formal bounds on watermark contraction.
- We conduct extensive experiments demonstrating that DiffErase achieves strong removal performance across diverse audio domains and watermarking systems while preserving perceptual quality.

## 2. Related Work

**Neural audio watermarking.** Audio watermarking embeds a hidden signal into an audio carrier for downstream attribution, such as copyright verification and provenance tracking. A typical system consists of an *embedder* that inserts a watermark (optionally carrying a message) and a *detector* that verifies and/or recovers it. Practical methods must satisfy two key requirements: (i) *fidelity*, meaning the watermark is imperceptible to listeners, and (ii) *robustness*, meaning the watermark survives common manipulations.

Traditional audio watermarking relies on handcrafted embedding and detection rules, making it difficult to balance fidelity and robustness. Recent neural approaches adopt end-to-end training, jointly optimized embedding, detection, fidelity, and robustness. Existing methods mainly differ in their embedding strategies. AudioSeal (Roman et al., 2024) and SilentCipher (Singh et al., 2024) employ a generator–detector design and embed watermarks into learned representations. WavMark (Chen et al., 2023) and IDEAW (Li et al., 2024) use invertible neural networks to model watermark embedding and detection as reversible transformations. TimbreWM (Liu et al., 2023b) and DeAR (Liu et al., 2023a)

embed watermark information in the frequency domain. While these methods demonstrate strong robustness against common distortions, their vulnerability to diffusion-based removal attacks remains largely unexplored.

**Attacks on audio watermarking.** Existing attacks vary in their assumptions and effectiveness. Signal-level transformations (Wen et al., 2025), such as pitch shifting, filtering, or lossy compression, require no knowledge of the watermark scheme but often degrade audio quality when applied strongly enough to remove watermarks. Overwriting attacks (Yao et al., 2025) embed a second watermark to interfere with the original, but they typically assume knowledge of the target watermarking architecture. Adversarial attacks (Liu et al., 2024) formulate watermark removal as an optimization problem, iteratively crafting perturbations that fool the detector while preserving perceptual quality. However, they require query access to the detector, which may be impractical when the watermarking scheme is private. These limitations motivate our work: a removal attack that preserves perceptual quality under minimal assumptions.

**Diffusion models for audio.** Diffusion models (Ho et al., 2020; Song et al., 2020) define a forward noising process and a learned reverse denoising process that maps noise back to the data distribution. DiffWave (Kong et al., 2020b) applies this framework to generate high-quality audio from mel-spectrograms and is commonly used as a neural vocoder. Conditional diffusion extends this paradigm to controlled generation, including text-to-audio (Liu et al., 2023c; Yang et al., 2023) and image-to-audio synthesis (Huang et al., 2023). Beyond generation, diffusion models have also been applied to audio restoration, reconstructing missing or corrupted content based on learned priors (Wang et al., 2023).

Our work builds on diffusion-based regeneration, which has been explored as both a defense and an attack strategy. As a defense, DiffPure (Nie et al., 2022) and its audio-domain variants (Wu et al., 2023; Guo et al., 2024) leverage diffusion denoising to suppress adversarial perturbations. As an attack, Zhao et al. (2024) demonstrate that regeneration can remove invisible image watermarks by projecting perturbed samples back onto the natural data manifold. These findings suggest diffusion regeneration can suppress low-magnitude, imperceptible signals. However, prior work focuses exclusively on the image domain; whether this vulnerability extends to audio watermarking remains unexplored.

## 3. Problem Formulation and Threat Model

### 3.1. Problem formulation

Let $x \in \mathbb{R}^T$ denote a clean audio waveform. An audio watermarking system consists of an embedding algorithm and a detection (or extraction) algorithm.

**Embedding.** Given an audio signal $x$, an optional message $m$, and a secret key $k \in \mathcal{K}$ (equivalently, secret model parameters), the embedder outputs a watermarked waveform:

$$x_w = \mathsf{Embed}(x, m, k) \in \mathbb{R}^T. \tag{1}$$

For message-based schemes, $m \in \{0,1\}^L$ is an $L$-bit payload. Presence-only schemes embed a zero-bit watermark for detection, without carrying an explicit message.

**Detection and extraction.** For message-based schemes, an extractor recovers the embedded message:

$$\widehat{m} = \mathsf{Ext}(x, k) \in \{0,1\}^L. \tag{2}$$

If $x$ is watermarked with payload $m$, then $\widehat{m}$ should match $m$; otherwise $\widehat{m}$ is a random bit string.

For presence-only schemes, a detector outputs a detection score indicating watermark presence:

$$s = \mathsf{Det}(x, k) \in [0,1]. \tag{3}$$

A binary decision is obtained by thresholding:

$$\mathsf{Result} = \mathbb{I}[s \geq \gamma] \in \{0,1\}, \tag{4}$$

where $\mathsf{Result} = 1$ indicates watermark presence, $\mathsf{Result} = 0$ indicates absence, and $\gamma$ is a detection threshold.

In this study, we evaluate our attack on five state-of-the-art open-source neural audio watermarking systems: AudioSeal (Roman et al., 2024), TimbreWM (Liu et al., 2023b), WavMark (Chen et al., 2023), Perth (Resemble AI, 2025), and SilentCipher (Singh et al., 2024).

### 3.2. Threat model

**Adversary's capabilities.** We consider a *black-box* adversary who: (i) has access only to the watermarked audio $x_w$; (ii) has no knowledge of the watermarking scheme (architecture, weights, or hyperparameters), and cannot query the detector or extractor; (iii) has sufficient computational resources to process $x_w$ using a diffusion model.

**Adversary's objective.** The adversary aims to produce a modified audio $\widehat{x}$ from $x_w$, disabling watermark verification while preserving perceptual quality. Formally, the attack succeeds if either

$$\mathsf{Result}(\widehat{x}, k) = 0 \quad \text{or} \quad \mathrm{Acc}\big(\mathsf{Ext}(\widehat{x}, k), m\big) < \eta, \tag{5}$$

where $\mathrm{Acc}(\cdot,\cdot)$ denotes bit accuracy and $\eta$ is a threshold below which message recovery is considered failed. Meanwhile, perceptual quality should be maintained:

$$\mathcal{Q}(\widehat{x}, x_w) \leq q_0, \tag{6}$$

where $\mathcal{Q}(\cdot,\cdot)$ is a perceptual distance metric and $q_0$ is a threshold beyond which degradation becomes perceptible.

# 4. DiffErase Attack

## 4.1. Preliminaries: audio diffusion models

Let $x_0 \in \mathbb{R}^d$ denote an audio representation. Diffusion models (Ho et al., 2020; Song et al., 2020) consist of (i) a fixed forward process that gradually adds noise and (ii) a learned reverse process that reconstructs data from noise.

**Forward process.** The forward process is a fixed Markov chain with variance schedule $\{\beta_t\}_{t=1}^N$. Let $\alpha_t \triangleq 1-\beta_t$ and $\bar{\alpha}_t \triangleq \prod_{s=1}^t \alpha_s$. The transition is

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}\, x_{t-1},\, \beta_t \mathbf{I}), \tag{7}$$

which yields a closed-form marginal distribution:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\, (1-\bar{\alpha}_t)\mathbf{I}\big). \tag{8}$$

Equivalently, $x_t$ can be sampled via reparameterization:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{9}$$

As $t$ increases, $x_t$ approaches an isotropic Gaussian.

**Reverse process.** The reverse process is modeled as a learned Markov chain:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t),\, \Sigma_t), \tag{10}$$

where $\Sigma_t$ is typically fixed. A standard parameterization predicts the forward noise using a neural network $\epsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t) \right). \tag{11}$$

The model is trained to minimize:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon, t}\big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\big], \tag{12}$$

## 4.2. DiffErase: diffusion-based watermark removal

We propose DiffErase, a diffusion-based attack that removes audio watermarks using a pre-trained diffusion model as a generative prior. Unlike standard diffusion-based generation, which starts from pure Gaussian noise, DiffErase follows an SDEdit-style procedure (Meng et al., 2021) with two stages: (i) *diffusion erasure*, which perturbs watermarked audio to an intermediate noise level, and (ii) *semantic reconstruction*, which applies the learned reverse dynamics to recover a clean signal. Crucially, DiffErase requires no knowledge of the target watermarking systems.

A key challenge is selecting an appropriate audio representation for diffusion. Waveform-level diffusion produces oversmoothed content and temporal drift. Linear spectrogram diffusion generates magnitude inconsistent with the original phase, producing audible artifacts(see Appendix C.4 for empirical comparisons). In contrast, mel-spectrograms capture salient structure such as energy contours and temporal envelopes while enabling high-quality reconstruction using modern vocoders, and we adopt this representation throughout. We model watermarked audio as $x_w = x_0 + \delta$, where $x_0$ is the clean audio and $\delta$ is the watermark perturbation.

**Phase I: diffusion erasure (forward noising).** We diffuse $x_w$ to an intermediate step $t^* \in \{1, \ldots, N\}$ using the closed-form forward marginal in (9):

$$x_{t^*} = \sqrt{\bar{\alpha}_{t^*}}\, x_w + \sqrt{1-\bar{\alpha}_{t^*}}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{13}$$

The hyperparameter $t^*$ controls a trade-off between watermark removal and reconstruction fidelity: a larger $t^*$ injects more noise, which better suppresses structured watermark signals but makes reverse reconstruction more challenging.

**Phase II: semantic reconstruction (reverse denoising).** Starting from $x_{t^*}$, we apply the reverse sampler induced by the pre-trained diffusion model from $t^*$ to 0:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z_t, \tag{14}$$

where $z_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We denote the final reconstruction by $\widehat{x}_0$. Since $\epsilon_\theta$ is trained on clean audio, the reverse dynamics tend to move samples toward high-density regions of the clean data distribution and suppress off-distribution perturbations introduced by watermarking. The complete procedure can be written as

$$\textsc{DiffErase}(x_w, t^*) \triangleq \textsc{Rev}(x_{t^*};\, t^* \to 0), \tag{15}$$

where $\textsc{Rev}(\cdot;\, t^* \to 0)$ denotes reverse sampling as in (14).

**Instantiations.** We implement two variants of DiffErase, both operating on mel-spectrograms:

(i) *Mel-spectrogram diffusion.* We apply DiffErase directly in the mel-spectrogram domain and reconstruct the waveform with a neural vocoder:

$$\widehat{x}_0 = \textsc{Voc}\big(\textsc{DiffErase}(\textsc{Mel}(x_w), t^*)\big), \tag{16}$$

where $\textsc{Mel}(\cdot)$ converts a waveform to a mel-spectrogram and $\textsc{Voc}(\cdot)$ inverts it back to a waveform.

(ii) *Latent diffusion.* Following Rombach et al. (2022), we encode mel-spectrograms into a learned latent space using a pretrained variational autoencoder $(\textsc{Enc}, \textsc{Dec})$, apply $\textsc{DiffErase}$ in the latent space, and decode back:

$$\widehat{x}_0 = \textsc{Voc}\big(\textsc{Dec}(\textsc{DiffErase}(\textsc{Enc}(\textsc{Mel}(x_w)), t^*))\big). \tag{17}$$

Latent diffusion reduces computational cost while achieving comparable attack performance. We compare these two variants in Section 5.1.

### 4.3. Theoretical analysis

We provide a theoretical justification for DiffErase. We model an imperceptible watermark as a small perturbation that moves the signal off the clean-audio manifold, and show that the diffusion prior suppresses these off-manifold components along the reverse trajectory.

**Definition 4.1** ($\Delta$-imperceptibility)**.** Let clean audio $x \in \mathbb{R}^d$ lie on a low-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$. Watermarked audio is constructed as $x_w = x + \delta$, where the perturbation satisfies $\|\delta\|_2 \le \Delta$ for a small constant $\Delta > 0$ to ensure imperceptibility.

**Definition 4.2** ($\tau$-margin detection)**.** Let $S(\cdot; k) \in \mathbb{R}$ denote a key-dependent detection statistic with key $k$. A watermark is *detected* if $S(x_w; k) \ge \tau$ for threshold $\tau > 0$. For processed audio $\widehat{x}$ derived from a watermarked input, we say the watermark is *removed* if $S(\widehat{x}; k) < \tau$.

**Off-manifold structure.** Following Gilmer et al. (2018); Stutz et al. (2019), small perturbations that do not alter semantics tend to be orthogonal to the data manifold's tangent space $\mathcal{T}_x\mathcal{M}$. While watermarks contain structure to enable decoding, they are designed to be imperceptible and statistically distinct from the natural audio distribution. We therefore decompose the watermark perturbation $\delta = \delta_{\parallel} + \delta_{\perp}$, where $\delta_{\parallel} \in \mathcal{T}_x\mathcal{M}$ and $\delta_{\perp} \perp \mathcal{T}_x\mathcal{M}$, and assume the off-manifold component dominates, i.e., $\|\delta_{\perp}\|_2 \gg \|\delta_{\parallel}\|_2$.

**Diffusion dynamics.** We analyze the attack process by coupling the watermarked trajectory with a reference clean trajectory. The forward phase diffuses $x_w$ to an intermediate timestep $t^*$, which scales the signal by $\sqrt{\bar{\alpha}_{t^*}}$ and injects Gaussian noise. The reverse phase follows the deterministic dynamics given by the probability flow ODE with score model $s_\theta(x, t) \approx \nabla_x \log p_t(x)$, where $p_t$ is the marginal distribution at diffusion time $t$.

**Lemma 4.3** (Score restores off-manifold deviations)**.** *Assume the manifold hypothesis and a local Gaussian approximation of $p_t$, the score function points towards the high-density region. For a watermarked state $x_t$ with off-manifold component $\Pi_\perp(x_t)\delta$, there exists $c_t > 0$ such that*

$$\langle s_\theta(x_t, t), \Pi_\perp(x_t)\delta \rangle \le -c_t \|\Pi_\perp(x_t)\delta\|_2^2, \tag{18}$$

*Proof sketch.* Since $p_t$ concentrates near $\mathcal{M}$, the marginal distribution in the normal direction approximates a Gaussian centered on the manifold. Consequently, the score function acts as a linear restoring force opposing off-manifold deviations. A detailed proof is provided in Appendix A.1.

**Lemma 4.4** (One-step contraction of watermark residue)**.** *Let $x_t$ and $x_t^{\text{clean}}$ denote two coupled reverse trajectories initialized from the watermarked and clean states, respectively. Define the watermark residue at time $t$ as $r_t \triangleq x_t - x_t^{\text{clean}}$. Under Lemma 4.3 and the assumption that residue is dominated by its off-manifold component, there exists a contraction factor $\rho_t \in (0, 1)$ such that*

$$\|r_{t-1}\|_2 \le \rho_t \|r_t\|_2. \tag{19}$$

*Proof sketch.* Consider one reverse step for both trajectories and subtract them to obtain an update rule for $r_t$. The score difference contributes a drift opposite to the normal component of $r_t$ (Lemma 4.3). Since the off-manifold component dominates, this restoring effect yields the contraction in Eq. (19). A detailed derivation is provided in Appendix A.2.

**Theorem 4.5** (Exponential decay of watermark residue)**.** *By combining the forward noising at timestep $t^*$ with the one-step contraction in Lemma 4.4, the final residue after reverse reconstruction satisfies*

$$\|r_0\|_2 \le \underbrace{\sqrt{\bar{\alpha}_{t^*}}}_{\text{forward scaling}} \cdot \underbrace{\left(\prod_{t=1}^{t^*} \rho_t\right)}_{\text{reverse contraction}} \cdot \Delta. \tag{20}$$

*Moreover, for any detection threshold $\tau > 0$, there exists a minimum diffusion steps $t^*_{\min}$ such that for all $t^* > t^*_{\min}$, the watermark becomes undetectable ($S(\widehat{x_0}; k) < \tau$).*

Theorem 4.5 shows that DiffErase suppresses watermark residue via two complementary mechanisms: (i) forward noising scales the signal component by ($\sqrt{\bar{\alpha}_{t^*}}$), attenuating the initial watermark perturbation, and (ii) reverse reconstruction yields geometric contraction with factors ($\rho_t$), filtering out off-manifold components. The full proof and derivation of $t^*_{\min}$ are provided in Appendix A.3.

## 5. Evaluation

**Setup.** We implement DiffErase with two diffusion backbones: (i) mel-spectrogram diffusion (von Platen et al., 2022) with BigVGAN (Lee et al., 2023) as the vocoder, and (ii) latent diffusion on mel-spectrograms (Liu et al., 2023c) with HiFi-GAN (Kong et al., 2020a) as the vocoder. Note that the diffusion process operates on integer timesteps $t \in \{1, \ldots, N\}$ ($N = 1000$ in our implementation), we report the noise level as a normalized ratio $t^* = t/N \in (0, 1]$ for clarity. Training configurations and implementation details are provided in Appendix B.1.

**Datasets.** We evaluate DiffErase across three audio domains: speech, music, and environmental sounds. For speech, we use the *100-hour* subset of LibriSpeech (Panayotov et al., 2015). For music, we use a subset of *FMA-small* from the Free Music Archive (FMA) (Defferrard et al., 2016). For environmental sounds, we use Clotho (Drossos et al., 2020). We randomly sample 100 clips from each domain for evaluation (more details refer to Appendix B.3).

*Table 1.* **Comparison with baselines** on the speech domain. Left: audio quality metrics (higher is better). Right: watermark detection measured by TPR@1%FPR (lower is better); ✗ indicates TPR $< 0.01$.

| Type | Attack | Audio Quality | | | Watermark Detection(TPR@1%FPR ↓) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SQUIM-MOS↑ | ViSQOL↑ | MUSHRA↑ | AudioSeal | WavMark | TimbreWM | Perth | SilentCipher |
| **Signal-level** | Pitch shift | 4.054 | 1.165 | 61.66 | ✗ | ✗ | ✗ | ✗ | ✗ |
| | Time stretch | 4.072 | 1.502 | 66.25 | 1.00 | 0.95 | 1.00 | 1.00 | ✗ |
| | Low-pass filter | 3.807 | 3.214 | 91.73 | 1.00 | 1.00 | 1.00 | 1.00 | 0.50 |
| | High-pass filter | 2.757 | 1.579 | 73.20 | 1.00 | 1.00 | 1.00 | 1.00 | 0.53 |
| | Additive noise | 3.062 | 1.063 | 25.64 | ✗ | ✗ | ✗ | ✗ | ✗ |
| **Codec** | MP3 | 4.503 | 4.123 | 96.42 | 1.00 | 0.97 | 1.00 | 1.00 | 0.34 |
| | EnCodec | 4.369 | 3.708 | 96.97 | 1.00 | ✗ | ✗ | 0.50 | ✗ |
| **Adaptive** | Square Attack | 3.025 | 2.567 | 54.07 | ✗ | ✗ | 0.28 | ✗ | ✗ |
| | **DIFFERASE-LATENT** | 4.214 | 3.477 | 87.73 | ✗ | ✗ | ✗ | ✗ | ✗ |
| | **DIFFERASE-MEL** | 4.423 | 3.961 | 93.81 | ✗ | ✗ | ✗ | ✗ | ✗ |

**Target watermarking systems.** We evaluate DiffErase against five state-of-the-art neural audio watermarking systems: AudioSeal (Roman et al., 2024), TimbreWM (Liu et al., 2023b), WavMark (Chen et al., 2023), Perth (Resemble AI, 2025), and SilentCipher (Singh et al., 2024). All systems are configured according to their official releases.

**Attack baselines.** We compare DiffErase against three categories of removal attacks. (i) *signal-level attacks*, like pitch shifting, time stretching, filtering, and additive Gaussian noise; (ii) *codec-based attacks*, including traditional codecs (e.g., MP3) and neural codecs (e.g., EnCodec); (iii) *adaptive attacks*, specifically Square Attack (Andriushchenko et al., 2020), a query-based adversarial attack. Details of all attack baselines are provided in Appendix B.2.

**Evaluation metrics.** We follow the evaluation protocol of O'Reilly et al. (2025). For watermark removal, we report the true positive rate at a fixed false positive rate, denoted as TPR@1%FPR; lower values indicate stronger removal. To assess perceptual quality after attack, we use both objective and subjective metrics: (1) SQUIM-MOS (Kumar et al., 2023), a non-intrusive metric estimating mean opinion score (MOS) on a 1–5 scale without reference audio; (2) ViSQOL (Chinen et al., 2020), which measures spectro-temporal similarity between reference and test audio on a 1–5 scale; and (3) MUSHRA, a subjective listening test where 16 participants rate samples on a 0–100 scale (details in Appendix B.4). Since attackers have access only to the watermarked audio $x_w$, we use $x_w$ as the reference for computing perceptual metrics.

### 5.1. Results and analysis

**Watermark removal and quality preservation.** Table 1 presents results on speech domain. Signal-level transformations generally fail to remove watermarks without causing severe quality degradation. Pitch shifting disables all watermark detectors but severely alters content and timbre, resulting in a low ViSQOL of 1.165 and MUSHRA of 61.66. Additive noise also removes all watermarks but introduces noticeable noise (MUSHRA:25.64). Other signal-level attacks, such as time stretching and frequency filtering, are largely ineffective, as these watermarking systems are typically trained to withstand such distortions.

*Table 2.* **DIFFERASE-MEL performance across domains.** Each entry shows values before/after attack. MUSHRA (higher is better) measures subjective audio quality; TPR@1%FPR (lower is better) measures watermark detectability.

| Domain | System | MUSHRA↑ | TPR@1%FPR ↓ |
|---|---|---|---|
| Speech | AudioSeal | 95.31 / 93.19 | 1.00 / 0.00 |
| | WavMark | 98.38 / 96.12 | 1.00 / 0.00 |
| | TimbreWM | 95.62 / 95.06 | 1.00 / 0.00 |
| | Perth | 92.31 / 90.69 | 1.00 / 0.00 |
| | SilentCipher | 96.69 / 94.00 | 1.00 / 0.00 |
| Music | AudioSeal | 95.62 / 87.12 | 1.00 / 0.00 |
| | WavMark | 92.75 / 85.31 | 1.00 / 0.00 |
| | TimbreWM | 95.00 / 84.06 | 1.00 / 0.01 |
| | Perth | 92.62 / 84.94 | 1.00 / 0.46 |
| | SilentCipher | 93.31 / 90.12 | 1.00 / 0.00 |
| Env. | AudioSeal | 92.00 / 83.62 | 1.00 / 0.00 |
| | WavMark | 94.06 / 87.94 | 1.00 / 0.00 |
| | TimbreWM | 93.62 / 86.38 | 0.97 / 0.00 |
| | Perth | 90.69 / 85.25 | 1.00 / 0.19 |
| | SilentCipher | 94.88 / 89.19 | 1.00 / 0.00 |

Codec-based attacks demonstrate stronger performance. Most watermarking methods remain robust to MP3 compression, maintaining TPR@1%FPR $\approx 1.00$. EnCodec achieves partial success against WavMark, TimbreWM, and SilentCipher, but fails against AudioSeal and Perth. Notably, both codecs preserve high perceptual quality (ViSQOL $> 3.7$ and MUSHRA $> 96$).

Adversarial attacks such as Square Attack degrade watermark detection by optimizing against detector outputs. However, they also introduce noticeable artifacts, yielding a MUSHRA score of only 54.07. In contrast, both DiffErase

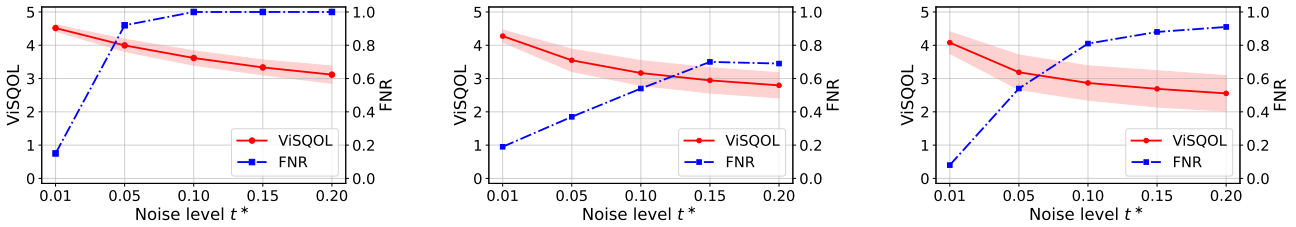



*Figure 2.* **Effect of noise level $t^*$ for DIFFERASE-MEL.** Trade-off between audio quality (ViSQOL, left axis) and watermark removal (FNR $= 1 -$ TPR, right axis), evaluated on Perth. **Left:** Speech. **Middle:** Music. **Right:** Environment.

variants suppress watermark detection to TPR $= 0$ across all tested watermarking schemes while preserving perceptual quality. At noise level $t^* = 0.1$, DIFFERASE-LATENT achieves SQUIM-MOS of 4.214 and ViSQOL of 3.477, and DIFFERASE-MEL achieves SQUIM-MOS of 4.423 and ViSQOL of 3.961. Subjective MUSHRA scores further confirm that DiffErase maintains audio quality better than competing attacks. More comparison results on music and environmental sound are provided in Appendix C.1.

Table 2 reports the overall performance of DIFFERASE-MEL across all three domains. DiffErase consistently reduces TPR@1%FPR from 1.00 to near 0.00 for most watermarking systems, with limited perceptual degradation (MUSHRA drops of 2–10 points). The commercial watermark Perth exhibits stronger robustness: while DiffErase fully removes Perth watermarks on speech (TPR $= 0.00$), residual detection remains on music (TPR $= 0.46$) and environmental sounds (TPR $= 0.19$). We analyze watermark strength by measuring the $\ell_2$ distance between clean and watermarked audio. As shown in Figure 3, Perth induces substantially larger perturbations than all other watermarking schemes, approximately 4–10$\times$ higher, which explains its robustness to DiffErase and its impact on perceptual quality.

**Effect of noise level between watermark removal and audio quality.** We evaluate how the noise level $t^*$ controls the trade-off between watermark removal and perceptual quality. As illustrated in Section 4.3 (Theorem 4.5), a threshold $t^*_{\min}$ exists above which imperceptible watermarks can be removed. Larger $t^*$ provides stronger denoising suppression of the embedded signal, improving removal but degrading fidelity. Smaller $t^*$ preserves perceptual quality but cannot effectively reduce watermark. We use Perth as the target since it is the most robust watermark in our evaluation and induces the largest perturbation, as shown in Figure 3. Result are reproted for DIFFERASE-MEL; results of DIFFERASE-LATENT are provided in Appendix C.2.

Figure 2 plots ViSQOL (left axis) and FNR $= 1 -$ TPR (right axis) as $t^*$ increases from 0.01 to 0.2. Across all domains, increasing $t^*$ improves watermark removal (higher FNR) while gradually reducing audio quality (lower ViSQOL). On speech, Perth becomes undetectable at $t^* \geq 0.10$ (FNR $\approx 1$) while quality remains high (ViSQOL $> 3.5$). On music and environmental sounds, FNR increases substantially with $t^*$ but cannot be fully removed even at $t^* = 0.20$. This is consistent with our earlier observation that Perth embeds a stronger watermark signal, which degrades perceptual quality but enhances robustness. It requires a higher diffusion noise level to completely eliminate.

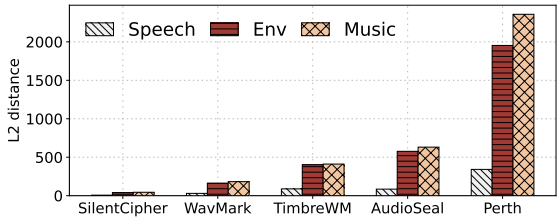



*Figure 3.* **The $\ell_2$ distance between clean and watermarked audio** across five watermarking methods on three domains. Perth embeds substantially stronger perturbations.

**Spectrogram visualization.** Figure 4 visualizes spectrograms for five audio examples processed by DIFFERASE-MEL at $t^* = 0.1$. The top row shows original (clean) audio, the middle row shows watermarked samples, and the bottom row shows DiffErase outputs. Different watermarking schemes embed watermarks into different time–frequency regions (highlighted by yellow boxes). After DiffErase, these structured patterns are visibly attenuated or removed.

The watermark patterns vary across schemes. AudioSeal, TimbreWM, and Perth embed repeated patterns to ensure robustness, introducing visible band-like structures. These structures become much less pronounced after DiffErase. WavMark embeds small distortions in low-energy regions, making the watermark less perceptible but still detectable. DiffErase disrupts these localized patterns as well. SilentCipher shows only subtle and spatially spread perturbations with less obvious spectrogram patterns, which explains its lower robustness in our evalution. Across all samples, DiffErase eliminates watermark patterns while preserving the main acoustic structure, though some fine-grained details become slightly smoothed. This smoothing is controlled by $t^*$: smaller $t^*$ values better preserve details but may leave

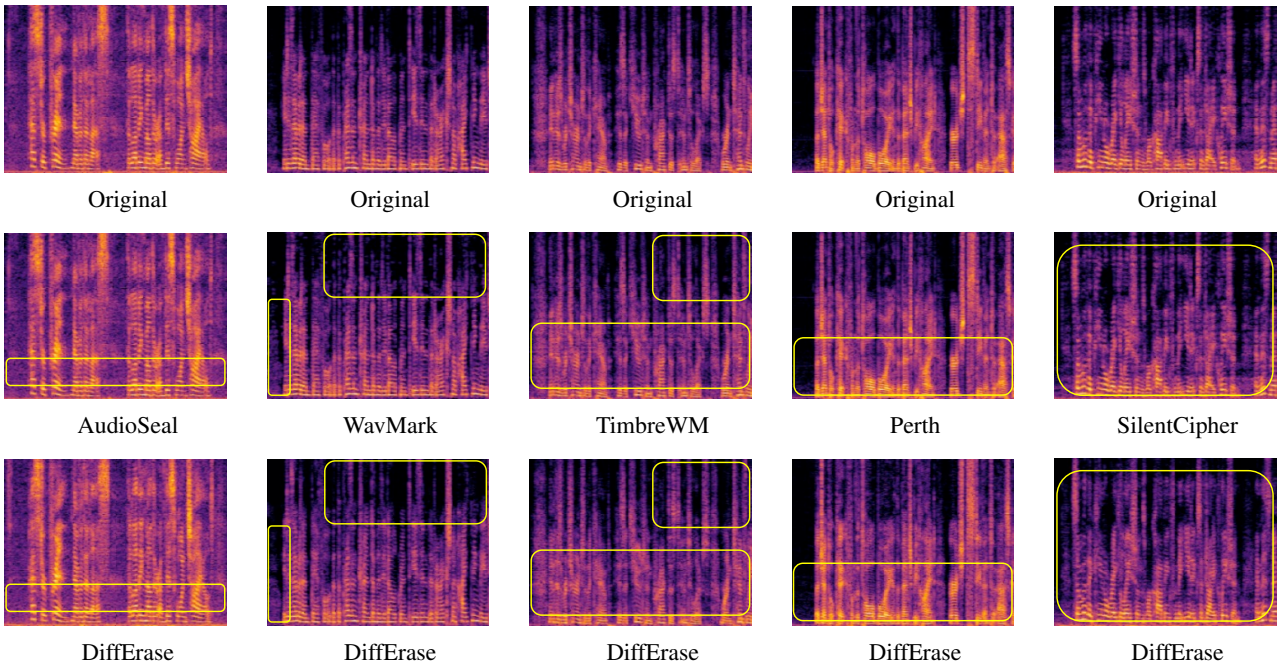



*Figure 4.* **Spectrogram visualization.** Top: original audio. Middle: watermarked audio. Bottom: after DIFFERASE-MEL ($t^* = 0.1$). Watermark patterns (middle row) are attenuated while acoustic content is preserved.

residual watermark evidence, while larger values improve removal at the cost of additional smoothing.

*Table 3.* **Ablation study** on speech. Full results on music and environment are provided in Appendix C.3.

| Method | TimbreWM | Perth | WavMark |
|---|---|---|---|
| **(i) mel-to-waveform** | | | |
| Griffin-Lim algorithm (GLA) only | 1.00 | 1.00 | 1.00 |
| DIFFERASE-MEL + GLA | 0.00 | 0.18 | 0.00 |
| DIFFERASE-LATENT + GLA | 0.00 | 0.00 | 0.00 |
| **(ii) diffusion sampler** | | | |
| DIFFERASE-MEL (DDPM) | 0.00 | 0.00 | 0.00 |
| DIFFERASE-MEL (DDIM) | 0.00 | 0.04 | 0.00 |
| DIFFERASE-LATENT (DDPM) | 0.00 | 0.00 | 0.00 |
| DIFFERASE-LATENT (DDIM) | 0.00 | 0.02 | 0.00 |

**Ablation study.** We conduct ablations to verify that watermark removal is primarily driven by the diffusion process rather than other pipeline components. We evaluate on three robust watermarking systems (TimbreWM, Perth, and WavMark), excluding AudioSeal and SilentCipher because their detection fails under waveform→spectrogram→waveform conversion alone.

To isolate the contribution of DiffErase, we reconstruct waveforms using the Griffin–Lim algorithm (GLA) (Griffin & Lim, 1984), which introduces minimal spectral distortion. As shown in Table 3, Griffin–Lim reconstruction alone does not affect watermark detection (TPR@1%FPR $= 1.00$ for all systems). In contrast, adding diffusion perturbations and denoising at $t^* = 0.1$ substantially degrades detection: DIFFERASE-MEL reduces TimbreWM and WavMark to TPR $= 0.00$ and Perth to 0.18, while DIFFERASE-LATENT removes all watermarks completely. The stronger removal of DIFFERASE-LATENT is consistent with the additional information bottleneck introduced from the VAE encoder.

We also compare diffusion samplers. At the same noise level $t^* = 0.1$, DDPM shows more effective removal than DDIM (50-step schedule), even though DDIM provides a faster processing speed. This suggests that the fine-grained denoising trajectory better suppresses watermark residues.

## 6. Conclusion

We propose DiffErase, a black-box attack that removes audio watermarks by leveraging diffusion models as generative priors. Unlike existing attacks, DiffErase requires neither detector queries nor knowledge of the watermarking schemes. Our theoretical analysis shows that diffusion dynamics suppress watermarks by contracting off-manifold perturbations along the reverse trajectory. Extensive experiments across three audio domains and five state-of-the-art watermarking systems demonstrate that DiffErase consistently removes watermarks while preserving perceptual quality. Our findings reveal a fundamental vulnerability in current audio watermarking designs: imperceptibility, while essential for practical deployment, inherently limits robustness against diffusion-based regeneration. This highlights the need for future watermarking designs to explicitly account for diffusion-based threats.

## Impact Statement

This paper studies the vulnerability of neural audio watermarking by presenting a black-box removal attack based on diffusion regeneration. The positive impact is to support more realistic robustness evaluation and facilitate stronger watermark designs. However, the proposed attack could potentially be misused to destroy provenance tracking or copyright protection. We are committed to responsible disclosure and encourage deployment-side mitigations and benchmarking protocols that account for diffusion-based threats.

## References


Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

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

Chinen, M., Lim, F. S., Skoglund, J., Gureev, N., O'Gorman, F., and Hines, A. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2020.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

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

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Guo, H., Wang, G., Chen, B., Wang, Y., Zhang, X., Chen, X., Yan, Q., and Xiao, L. Wavepurifier: Purifying audio adversarial examples via hierarchical diffusion models. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 1268–1282, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pp. 13916–13932. PMLR, 2023.

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

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020b.

Kumar, A., Tan, K., Ni, Z., Manocha, P., Zhang, X., Henderson, E., and Xu, B. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Lee, S. G., Ping, W., Ginsburg, B., Catanzaro, B., and Yoon, S. Bigvgan: A universal neural vocoder with large-scale training. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Li, K., Hu, X., Grishchenko, I., and Lie, D. Harmonicattack: An adaptive cross-domain audio watermark removal. *arXiv preprint arXiv:2511.21577*, 2025.

Li, P., Zhang, X., Xiao, J., and Wang, J. Ideaw: Robust neural audio watermarking with invertible dual-embedding. *arXiv preprint arXiv:2409.19627*, 2024.

Liu, C., Zhang, J., Fang, H., Ma, Z., Zhang, W., and Yu, N. Dear: A deep-learning-based audio re-recording resilient watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13201–13209, 2023a.

Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W., and Yu, N. Detecting voice cloning attacks via timbre watermarking. *arXiv preprint arXiv:2312.03410*, 2023b.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023c.

Liu, H., Guo, M., Jiang, Z., Wang, L., and Gong, N. Audiomarkbench: Benchmarking robustness of audio watermarking. *Advances in Neural Information Processing Systems*, 37:52241–52265, 2024.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

O'Reilly, P., Jin, Z., Su, J., and Pardo, B. Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech. *arXiv preprint arXiv:2504.10782*, 2025.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Resemble AI. Perth: Open audio watermarking tool. GitHub repository, 2025. URL https://github.com/resemble-ai/Perth.

Roman, R. S., Fernandez, P., Défossez, A., Furon, T., Tran, T., and Elsahar, H. Proactive detection of voice cloning with localized watermarking. *arXiv preprint arXiv:2401.17264*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., and Feizi, S. Robustness of AI-image detectors: Fundamental limits and practical attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dLoAdIKENc.

Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1), 2018.

Singh, M. K., Takahashi, N., Liao, W., and Mitsufuji, Y. Silentcipher: Deep audio watermarking. *arXiv preprint arXiv:2406.03822*, 2024.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Stutz, D., Hein, M., and Schiele, B. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6976–6987, 2019.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wang, Y., Ju, Z., Tan, X., He, L., Wu, Z., Bian, J., et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Wen, Y., Innuganti, A., Ramos, A. B., Guo, H., and Yan, Q. Sok: How robust is audio watermarking in generative ai models? *arXiv preprint arXiv:2503.19176*, 2025.

Wu, S., Wang, J., Ping, W., Nie, W., and Xiao, C. Defending against adversarial audio via diffusion model. *arXiv preprint arXiv:2303.01507*, 2023.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

Yao, L., Huang, C., Wang, S., Xue, J., Guo, H., Liu, J., Lin, P., Ohtsuki, T., and Pan, M. Yours or mine? overwriting attacks against neural audio watermarking. *arXiv preprint arXiv:2509.05835*, 2025.

Yoon, J., Hwang, S. J., and Lee, J. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pp. 12062–12072. PMLR, 2021.

Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., Vigna, G., Wang, Y.-X., and Li, L. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024.

# A. Proofs in Section 4.3

## A.1. Proof of Lemma 4.3

**Lemma A.1** (Score restores off-manifold deviations)**.** *Assume the manifold hypothesis and a local Gaussian approximation of $p_t$, the score function points towards the high-density region. For a watermarked state $x_t$ with off-manifold component $\Pi_\perp(x_t)\delta$, there exists $c_t > 0$ such that*

$$\langle s_\theta(x_t, t), \Pi_\perp(x_t)\delta\rangle \le -c_t\|\Pi_\perp(x_t)\delta\|_2^2, \tag{21}$$

*Proof.* We rely on the geometric interpretation of the score function established in prior works (Yoon et al., 2021; Chung et al., 2022). Under the manifold hypothesis, the diffusion marginal $p_t$ concentrates around $\mathcal{M}$ and can be locally approximated as a Gaussian centered at the manifold. As a result, for $x_t$ close to $\mathcal{M}$, the score has a normal component that points back toward the manifold. Concretely, we use the approximation

$$\Pi_\perp(x_t)s_\theta(x_t, t) \approx -\frac{1}{\sigma_t^2}(x_t - \Pi_{\mathcal{M}}(x_t)), \tag{22}$$

where $\Pi_{\mathcal{M}}(x_t)$ is the (local) projection of $x_t$ onto $\mathcal{M}$ and $\sigma_t$ is the noise scale at time $t$.

The watermark perturbation $\delta$ is modeled as predominantly off-manifold (as justified in Sec. 4.3), the deviation from the manifold is dominated by the watermark component: $x_t - \Pi_{\mathcal{M}}(x_t) \approx \Pi_\perp(x_t)\delta$. Substituting this into the inner product, we obtain:

$$\begin{aligned}
\langle s_\theta(x_t, t), \Pi_\perp(x_t)\delta\rangle &= \langle \Pi_\perp(x_t)s_\theta(x_t, t), \Pi_\perp(x_t)\delta\rangle && (23)\\
&\approx \left\langle -\frac{1}{\sigma_t^2}\Pi_\perp(x_t)\delta, \Pi_\perp(x_t)\delta \right\rangle && (24)\\
&= -\frac{1}{\sigma_t^2}\|\Pi_\perp(x_t)\delta\|_2^2. && (25)
\end{aligned}$$

Letting $c_t = 1/\sigma_t^2 > 0$. Replacing the approximation by an inequality to account for local modeling error gives (21). □

## A.2. Proof of Lemma 4.4

**Lemma A.2** (One-step contraction of watermark residue)**.** *Let $x_t$ and $x_t^{\text{clean}}$ denote two coupled reverse trajectories initialized from the watermarked and clean states, respectively. Define the watermark residue at time $t$ as $r_t \triangleq x_t - x_t^{\text{clean}}$. Under Lemma 4.3 and the assumption that residue is dominated by its off-manifold component, there exists a contraction factor $\rho_t \in (0, 1)$ such that*

$$\|r_{t-1}\|_2 \le \rho_t\|r_t\|_2. \tag{26}$$

*Proof.* We analyze the deterministic probability flow ODE (as an idealized model for the reverse dynamics). Consider a discretized reverse step from time $t$ to $t-1$ using a first-order Euler solver:

$$x_{t-1} \approx x_t + \eta_t s_\theta(x_t, t), \tag{27}$$

where $\eta_t > 0$ is an effective step size determined by the specific scheduler. Applying to both the watermarked trajectory $x_t$ and the clean trajectory $x_t^{\text{clean}}$ using the same step size yields

$$\begin{aligned}
r_{t-1} &= x_{t-1} - x_{t-1}^{\text{clean}} && (28)\\
&\approx (x_t + \eta_t s_\theta(x_t, t)) - (x_t^{\text{clean}} + \eta_t s_\theta(x_t^{\text{clean}}, t)) && (29)\\
&= r_t + \eta_t \left(s_\theta(x_t, t) - s_\theta(x_t^{\text{clean}}, t)\right). && (30)
\end{aligned}$$

Next, we relate the score difference to the residue. Lemma 4.3 implies that, in normal directions near $\mathcal{M}$, the score acts as a restoring force. Under a first-order local linearization around the clean trajectory, we approximate

$$s_\theta(x_t, t) - s_\theta(x_t^{\text{clean}}, t) \approx -c_t(x_t - x_t^{\text{clean}}) = -c_t r_t, \tag{31}$$

where $c_t \approx 1/\sigma_t^2 > 0$ determines the strength of the restoring force derived in Lemma 4.3.

Substituting this relationship into Eq. (30):

$$r_{t-1} \approx r_t - \eta_t c_t r_t = (1 - \eta_t c_t) r_t. \tag{32}$$

Taking the Euclidean norm, we obtain:

$$\|r_{t-1}\|_2 \approx |1 - \eta_t c_t| \|r_t\|_2. \tag{33}$$

Define the decay factor $\rho_t = |1 - \eta_t c_t|$. In standard diffusion schedules, the step size $\eta_t$ satisfies $0 < \eta_t c_t < 1$ (typically $0 < \eta_t c_t < 2$). Therefore, there exists $\rho_t \in (0, 1)$ such that:

$$\|r_{t-1}\|_2 \leq \rho_t \|r_t\|_2. \tag{34}$$

□

### A.3. Proof of Theorem 4.5

**Theorem A.3** (Exponential decay of watermark residue)**.** *By combining the forward noising at timestep $t^*$ with the one-step contraction in Lemma 4.4, the final residue after reverse reconstruction satisfies*

$$\|r_0\|_2 \leq \underbrace{\sqrt{\bar{\alpha}_{t^*}}}_{\textit{Forward scaling}} \cdot \underbrace{\left( \prod_{t=1}^{t^*} \rho_t \right)}_{\textit{Reverse contraction}} \cdot \Delta. \tag{35}$$

*Moreover, for any detection threshold $\tau > 0$, there exists a minimum diffusion steps $t^*_{\min}$ such that for all $t^* > t^*_{\min}$, the watermark becomes undetectable ($S(\widehat{x}_0; k) < \tau$).*

*Proof.* We first analyze the residue at the attack timestep $t^*$. Let $x_0$ denote the clean audio and $x_0^w = x_0 + \delta$ the watermarked audio, with $\|\delta\|_2 \leq \Delta$. We adopt a coupled forward process where both trajectories share the same Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At timestep $t^*$,

$$x_{t^*} = \sqrt{\bar{\alpha}_{t^*}} x_0 + \sqrt{1 - \bar{\alpha}_{t^*}} \epsilon, \tag{36}$$

$$x^w_{t^*} = \sqrt{\bar{\alpha}_{t^*}} (x_0 + \delta) + \sqrt{1 - \bar{\alpha}_{t^*}} \epsilon. \tag{37}$$

Define the residue at $t^*$ as $r_{t^*} \triangleq x^w_{t^*} - x_{t^*}$. Then

$$r_{t^*} = \sqrt{\bar{\alpha}_{t^*}} \delta. \tag{38}$$

Consequently, the norm of the perturbation at $t^*$ is scaled by the signal decay factor:

$$\|r_{t^*}\|_2 = \sqrt{\bar{\alpha}_{t^*}} \, \|\delta\|_2 \leq \sqrt{\bar{\alpha}_{t^*}} \, \Delta. \tag{39}$$

Applying Lemma 4.4 iteratively from $t^*$ down to 0 yields

$$\|r_0\|_2 \leq \rho_1 \|r_1\|_2 \leq \cdots \leq \left( \prod_{j=1}^{t^*} \rho_j \right) \|r_{t^*}\|_2. \tag{40}$$

Substituting (39) gives

$$\|r_0\|_2 \leq \sqrt{\bar{\alpha}_{t^*}} \left( \prod_{j=1}^{t^*} \rho_j \right) \Delta. \tag{41}$$

□

**Bound on diffusion steps $t^*_{\min}$.** To derive the lower bound on $t^*$ required for watermark removal $S(\widehat{x}_0; k) < \tau$, we link the residue norm to the detection statistic. Assume the detection function $S(\cdot; k)$ is $L$-Lipschitz continuous. For clean audio (no watermark), $S(x^{\text{clean}}; k) \approx 0$. For the reconstructed output $\widehat{x}_0$, we have

$$\begin{aligned} S(\widehat{x}_0; k) &\le |S(\widehat{x}_0; k) - S(x^{\text{clean}}; k)| + S(x^{\text{clean}}; k) \\ &\le L\|\widehat{x}_0 - x^{\text{clean}}\|_2 + S(x^{\text{clean}}; k) \\ &\approx L\|r_0\|_2 \\ &\le L\Delta\sqrt{\bar{\alpha}_{t^*}} \left(\prod_{t=1}^{t^*} \rho_t\right), \end{aligned} \tag{42}$$

where the final inequality uses (41).

Let $\bar{\rho} \triangleq \max_{t\in\{1,\dots,t^*\}} \rho_t < 1$, so that $\prod_{t=1}^{t^*} \rho_t \le \bar{\rho}^{t^*}$. A sufficient condition for $S(\widehat{x}_0; k) < \tau$ is therefore

$$L\Delta\sqrt{\bar{\alpha}_{t^*}}\, \bar{\rho}^{t^*} < \tau. \tag{43}$$

Taking logarithms yields

$$\ln(L\Delta) + \frac{1}{2}\ln \bar{\alpha}_{t^*} + t^* \ln \bar{\rho} < \ln \tau. \tag{44}$$

Since $\bar{\alpha}_{t^*} \in (0, 1)$ and $\bar{\rho} \in (0, 1)$, their logarithms are negative. Rearranging gives the equivalent condition

$$t^*|\ln \bar{\rho}| + \tfrac{1}{2}|\ln \bar{\alpha}_{t^*}| > \ln\left(\frac{L\Delta}{\tau}\right). \tag{45}$$

Because $\bar{\alpha}_{t^*}$ decreases monotonically with $t^*$, both terms on the left-hand side increase with $t^*$. Hence there exists a minimal $t^*_{\min}$ satisfying (45), and any $t^* \ge t^*_{\min}$ is sufficient to ensure $S(\widehat{x}_0; k) < \tau$ under the stated assumptions.

## B. Implementation Details

### B.1. DiffErase instantiations

We implement DiffErase with two complementary backbones:

**DIFFERASE-MEL.** This variant is built on DIFFUSER library (von Platen et al., 2022) and performs diffusion directly in the mel-spectrogram domain. We use a UNet2DModel as the denoiser, treating the mel-spectrogram as a single-channel image. The mel-spectrogram is computed with 80 mel bins. For waveform reconstruction, we use BigVGAN (Lee et al., 2023) (`bigvgan_v2_22khz_80band_256x`) as the vocoder.

**DIFFERASE-LATENT.** This variant is built on the AudioLDM pipeline (Liu et al., 2023c). Mel-spectrograms (64 mel bins) are first encoded into a latent space using a pretrained AutoencoderKL with latent channel dimension of 8. Diffusion is then performed in the VAE latent space using a UNet with the following configuration: image size 64, base channels 128, 2 residual blocks per stage, channel multipliers $[1, 2, 3, 5]$, and attention at resolutions $\{8, 4, 2\}$. The reconstructed mel-spectrogram is converted back to waveform using HiFi-GAN (Kong et al., 2020a).

### B.2. Baseline atacks

**Signal-level attacks.** Following the setting in O'Reilly et al. (2025), pitch shifting uses a random shift in $[-1, 1]$ semitones, and time stretching applies a random speed factor in $[0.95, 1.05]$. For frequency filtering, we apply a low-pass filter with a 4000 Hz cutoff and a high-pass filter with a 500 Hz cutoff. Additive noise is Gaussian with standard deviation $\sigma = 0.01$.

**Codec-based attacks.** We evaluate a traditional codec and a neural-based codec. MP3 compression is performed using FFmpeg at 32 kbps. EnCodec (Défossez et al., 2022) uses 24 kbps bandwidth.

**Adaptive attack.** We implement Square attack (Andriushchenko et al., 2020) in the spectrogram domain following Liu et al. (2024). For speech, we use a query budget of 10,000 and perturbation bound $\epsilon = 0.02$. For music and environmental sounds, we found $\epsilon = 0.02$ insufficient for successful attacks and increased it to $\epsilon = 0.2$.

*Table 4.* **Comparison with baselines** on the music domain. Left: audio quality metrics (higher is better). Right: watermark detection measured by TPR@1%FPR (lower is better); ✗ indicates TPR $< 0.01$.

| Type | Attack | Audio Quality | | Watermark Detection(TPR@1%FPR ↓) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ViSQOL↑ | MUSHRA↑ | AudioSeal | WavMark | TimbreWM | Perth | SilentCipher |
| Signal-level | Pitch shift | 4.061 | 78.33 | ✗ | ✗ | ✗ | 0.10 | ✗ |
| | Time stretch | 4.109 | 80.16 | ✗ | 0.83 | 0.87 | 1.00 | ✗ |
| | Low-pass filter | 3.203 | 87.28 | 1.00 | 1.00 | 1.00 | 1.00 | 0.51 |
| | High-pass filter | 3.673 | 78.61 | 1.00 | 1.00 | 1.00 | 0.99 | 0.58 |
| | Additive noise | 2.778 | 25.40 | 0.43 | ✗ | 0.12 | 0.43 | ✗ |
| Codec | MP3 | 4.505 | 94.39 | 1.00 | 0.84 | 0.94 | 1.00 | 0.42 |
| | EnCodec | 4.321 | 88.66 | 0.99 | ✗ | ✗ | 0.89 | ✗ |
| Adaptive | Square Attack | 2.960 | 31.24 | 0.18 | ✗ | 0.36 | 0.21 | ✗ |
| | DIFFERASE-LATENT | 4.163 | 91.20 | ✗ | ✗ | 0.01 | 0.35 | ✗ |
| | DIFFERASE-MEL | 3.938 | 86.31 | ✗ | ✗ | 0.01 | 0.46 | ✗ |

*Table 5.* **Comparison with baselines** on the environment sound domain. Left: audio quality metrics (higher is better). Right: watermark detection measured by TPR@1%FPR (lower is better); ✗ indicates TPR $< 0.01$.

| Type | Attack | Audio Quality | | Watermark Detection(TPR@1%FPR ↓) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ViSQOL↑ | MUSHRA↑ | AudioSeal | WavMark | TimbreWM | Perth | SilentCipher |
| Signal-level | Pitch shift | 4.061 | 76.20 | ✗ | ✗ | ✗ | 0.03 | ✗ |
| | Time stretch | 4.347 | 87.06 | ✗ | 0.85 | 0.97 | 1.00 | ✗ |
| | Low-pass filter | 3.102 | 92.53 | 1.00 | 0.94 | 0.13 | 1.00 | 0.32 |
| | High-pass filter | 3.706 | 84.80 | 1.00 | 1.00 | 0.97 | 1.00 | 0.59 |
| | Additive noise | 1.443 | 31.43 | 0.34 | ✗ | 0.11 | 0.39 | ✗ |
| Codec | MP3 | 3.979 | 97.21 | 1.00 | 0.65 | 0.82 | 0.99 | 0.22 |
| | EnCodec | 4.334 | 94.81 | 0.96 | ✗ | ✗ | 0.90 | ✗ |
| Adaptive | Square Attack | 2.365 | 58.20 | 0.11 | ✗ | 0.06 | 0.38 | ✗ |
| | DIFFERASE-LATENT | 3.308 | 87.07 | ✗ | ✗ | ✗ | 0.23 | ✗ |
| | DIFFERASE-MEL | 3.952 | 86.48 | ✗ | ✗ | ✗ | 0.19 | ✗ |

### B.3. Dataset details

We evaluate DiffErase across three audio domains to assess generalization. For speech, we use LibriSpeech (Panayotov et al., 2015), an English corpus derived from audiobooks containing 1,000 hours of speech. We train on the *train-clean-100* subset (100 hours) and evaluate on 100 randomly sampled clips from *test-clean*. For environmental sounds, we use Clotho (Drossos et al., 2020), an audio captioning dataset. We discard the captions and reserve 100 samples for evaluation, using the reminder for training. For music, we use *FMA-small* (Defferrard et al., 2016) from the Free Music Archive, which contains songs across various genres. We use all samples for training except 100 samples for evaluation.

### B.4. Subjective listening test

We conduct a MUSHRA listening test following ITU-R BS.1534 to evaluate perceptual quality. We randomly sample 5 audio clips per domain and include all attack samples in each trial. 18 participants completed the study, filtered unreliable evaluations using low-quality anchor samples, and 16 valid participants were retained for analysis. Participants rate the quality of processed audio samples on a scale of 0 to 100, with watermarked audio provided as the reference. Evaluations are conducted using the open-source webMUSHRA platform (Schoeffler et al., 2018) locally.

## C. Additional experimental results

### C.1. Comparison with baselines on music and environmental sounds

Table 4 and Table 5 present comparison results on the music and environmental sound domains, respectively. The overall performance is consistent with the speech domain.

Signal-level attacks show limited effectiveness. Pitch shifting removes most watermarks but degrades quality (MUSHRA of 78.33 on music and 76.20 on environmental). Additive noise achieves partial removal but introduces severe distortions (MUSHRA of 25.40 on music and 31.43 on environment). Time stretching and frequency filtering remains largely ineffective.

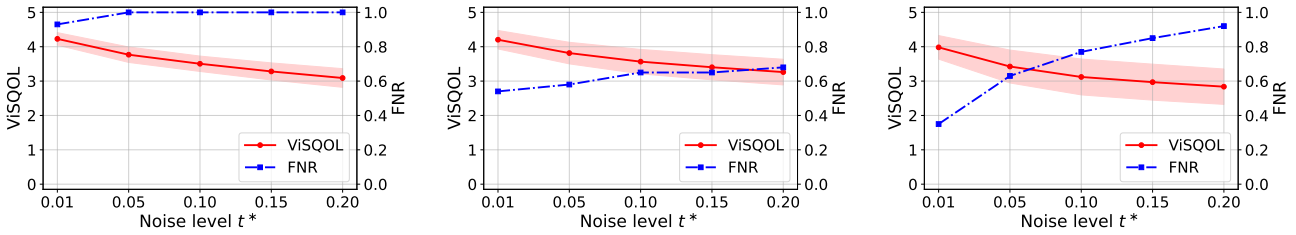



*Figure 5.* **Effect of noise level $t^*$ for DIFFERASE-LATENT.** Trade-off between audio quality (ViSQOL, left axis) and watermark removal (FNR $= 1 - \mathrm{TPR}$, right axis), evaluated on Perth. **Left:** Speech. **Middle:** Music. **Right:** Environment.

*Table 6.* **Ablation study** on music.

| Method | TimbreWM | Perth | WavMark |
|---|---|---|---|
| **(i) mel-to-waveform** | | | |
| Griffin-Lim algorithm (GLA) only | 1.00 | 1.00 | 1.00 |
| DIFFERASE-MEL + GLA | 0.00 | 0.46 | 0.00 |
| DIFFERASE-LATENT + GLA | 0.00 | 0.00 | 0.00 |
| **(ii) diffusion sampler** | | | |
| DIFFERASE-MEL (DDPM) | 0.01 | 0.46 | 0.00 |
| DIFFERASE-MEL (DDIM) | 0.02 | 0.75 | 0.00 |
| DIFFERASE-LATENT (DDPM) | 0.01 | 0.35 | 0.00 |
| DIFFERASE-LATENT (DDIM) | 0.02 | 0.44 | 0.00 |

*Table 7.* **Ablation study** on environment sound.

| Method | TimbreWM | Perth | WavMark |
|---|---|---|---|
| **(i) mel-to-waveform** | | | |
| Griffin-Lim algorithm (GLA) only | 0.89 | 1.00 | 1.00 |
| DIFFERASE-MEL + GLA | 0.00 | 0.58 | 0.00 |
| DIFFERASE-LATENT + GLA | 0.00 | 0.00 | 0.00 |
| **(ii) diffusion sampler** | | | |
| DIFFERASE-MEL (DDPM) | 0.00 | 0.19 | 0.00 |
| DIFFERASE-MEL (DDIM) | 0.00 | 0.34 | 0.00 |
| DIFFERASE-LATENT (DDPM) | 0.00 | 0.23 | 0.00 |
| DIFFERASE-LATENT (DDIM) | 0.00 | 0.23 | 0.00 |

Codec-based attacks preserve high quality, but fail to remove watermarks. MP3 compression maintains MUSHRA above 94 but fails to remove most watermarks. EnCodec successfully removes WavMark, TimbreWM, and SilentCipher while preserving audio quality, but fails against AudioSeal and Perth.

Even though Square attack shows relatively good removal performance on speech, it fails on music and environmental sounds despite increasing the perturbation bound to $\epsilon = 0.2$, which causes substantial quality degradation (MUSHRA of 31.24 on music and 58.20 on environment).

DiffErase achieves strong watermark removal across both domains while preserving competitive quality. On music, DIFFERASE-LATENT achieves ViSQOL of 4.16 and MUSHRA of 91.20, removing all watermarks except Perth ($\mathrm{TPR} = 0.35$). On environmental sounds, DIFFERASE-MEL achieves ViSQOL of 3.95 and MUSHRA of 86.48. Perth remains partially detectable owing to its stronger embedded perturbations observed in Figure 3.

### C.2. Noise level trade-off for DIFFERASE-LATENT

Figure 5 shows the quality-removal trade-off for DIFFERASE-LATENT. The trends are similar: increasing $t^*$ improves watermark removal (higher FNR) while reducing audio quality (lower ViSQOL). On speech, Perth becomes undetectable at $t^* \geq 0.05$. On music and environmental sounds, complete removal requires higher noise levels. Compared to DIFFERASE-MEL, DIFFERASE-LATENT achieves comparable removal at similar noise levels but with slightly lower ViSQOL, likely due to the additional compression from the VAE encoder.

### C.3. Ablation study on music and environmental sounds

Tables 6 and 7 extend the ablation study to the music and environmental sound domains. Consistent with the speech results, Griffin–Lim reconstruction alone does not remove watermarks ($\mathrm{TPR} = 1.00$ for most systems). Adding diffusion substantially improves removal. DDPM consistently outperforms DDIM across both domains, particularly for Perth. On music, DDPM reduces Perth to $\mathrm{TPR} = 0.46$ while DDIM only achieves 0.75. This confirms that the finer-grained denoising trajectory of DDPM better suppresses watermark residues.

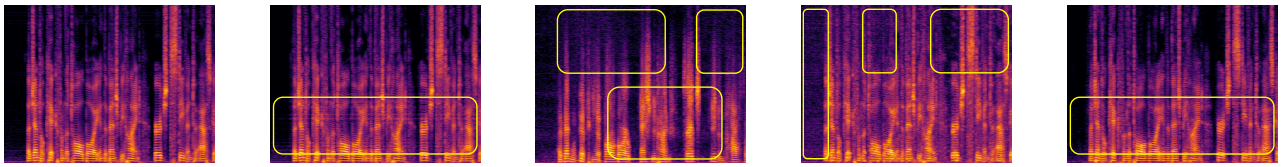
### C.4. Ablation on attack representation

We compare DiffErase operating on mel-spectrograms against alternative representations: (i) waveform using DiffWave (Kong et al., 2020b), and (ii) linear spectrogram with phase preservation following WavePurifier (Guo et al., 2024).

*Table 8.* **Ablation on audio representation.** We compare DiffErase with diffusion on different representations.

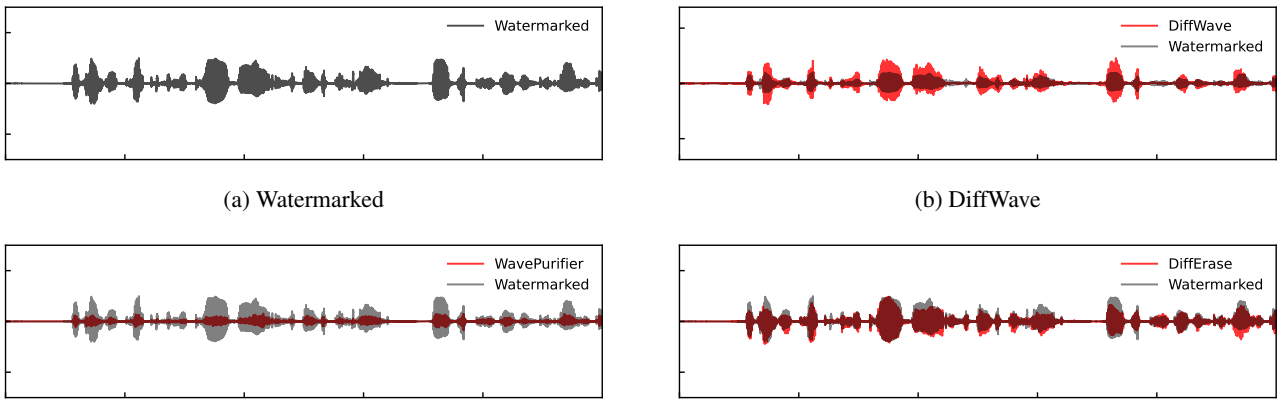
| Representation | Audio Quality | | | Watermark Detection(TPR@1%FPR ↓) | | | | |
|---|---|---|---|---|---|---|---|---|
| | SQUIM-MOS↑ | ViSQOL↑ | MUSHRA↑ | AudioSeal | WavMark | TimbreWM | Perth | SilentCipher |
| Waveform | 2.541 | 1.996 | 71.21 | ✗ | ✗ | ✗ | ✗ | ✗ |
| Linear spectrogram | 3.825 | 3.857 | 79.66 | 1.00 | 1.00 | 1.00 | 0.97 | 0.35 |
| Mel-spectrogram | 4.423 | 3.961 | 93.81 | ✗ | ✗ | ✗ | ✗ | ✗ |
| Mel-latent | 4.214 | 3.477 | 87.73 | ✗ | ✗ | ✗ | ✗ | ✗ |

As shown in Table 8, waveform diffusion successfully remove watermark but produces low-quality audio (SQUIM-MOS of 2.541, MUSHRA of 71.21). The high dimensionality of raw waveforms leads to over-smoothed content and temporal drift, failing to preserve perceptual details. This is evident in Figure 6(c), where harmonic structures are blurred, and in Figure 7(b), where the reconstructed waveform significantly deviates from the original envelope. Linear spectrogram diffusion improves quality but remains inferior to the mel-spectrogram. Its reconstruction relies on reusing the original phase, but the regenerated magnitude may be inconsistent with the preserved phase, producing audible artifacts, as shown in Figure 6(d) and Figure 7(c). In contrast, mel-spectrogram diffusion achieves the best trade-off: complete watermark removal while maintaining high audio quality. As shown in Figure 6(e), DiffErase preserves the main spectral structure while removing watermark perturbations, enabling stable diffusion and high-quality reconstruction via modern neural vocoders. This is further confirmed in Figure 7(d), where the reconstructed waveform closely follows the original temporal envelope.



(a) Original (b) Watermarked (c) DiffWave (d) WavePurifier (e) DiffErase

*Figure 6.* **Mel-spectrogram visualization** across different attack representations. DiffWave (c) over-smooths harmonic details. WavePurifier (d) introduces audible artifacts. DiffErase (e) preserves the spectral structure while removing watermark perturbations.




(a) Watermarked (b) DiffWave

(c) WavePurifier (d) DiffErase

*Figure 7.* **Waveform visualization** across different attack representations. Red waveforms show reconstructed audio, and gray shows input watermarked audio. DiffErase (d) closely matches the original temporal envelope.