

# HarmonicAttack: An Adaptive Cross-Domain Audio Watermark Removal

Kexin Li<sup>†</sup>, Xiao Hu<sup>†</sup>, Ilya Grishchenko, David Lie

University of Toronto

**Abstract**—The availability of high-quality, AI-generated audio raises security challenges such as misinformation campaigns and voice-cloning fraud. A key defense against the misuse of AI-generated audio is by watermarking it, so that it can be easily distinguished from genuine audio. As those seeking to misuse AI-generated audio may thus seek to remove audio watermarks, studying effective watermark removal techniques is critical to being able to objectively evaluate the robustness of audio watermarks against removal. Previous watermark removal schemes either assume impractical knowledge of the watermarks they are designed to remove or are computationally expensive, potentially generating a false sense of confidence in current watermark schemes.

We introduce HarmonicAttack, an efficient audio watermark removal method that only requires the basic ability to generate the watermarks from the targeted scheme and nothing else. With this, we are able to train a general watermark removal model that is able to remove the watermarks generated by the targeted scheme from any watermarked audio sample. HarmonicAttack employs a dual-path convolutional autoencoder that operates in both temporal and frequency domains, along with GAN-style training, to separate the watermark from the original audio. When evaluated against state-of-the-art watermark schemes AudioSeal, WavMark, and Silentcipher, HarmonicAttack demonstrates greater watermark removal ability than previous watermark removal methods with near real-time performance. Moreover, while HarmonicAttack requires training, we find that it is able to transfer to out-of-distribution samples with minimal degradation in performance.

## 1. Introduction

Recent advancements in generative audio models [1], [2] have revolutionized how we create, manipulate, and interact with sound. From text-to-speech systems and music generation to audio enhancement, these models demonstrate capabilities in producing realistic and high-quality audio.

The growing realism of synthetic audio has made it increasingly difficult to distinguish it from genuine audio. This indistinguishability, together with accessibility, poses significant risks of misinformation, impersonation, and fraud [3]. For instance, voice cloning enabled by AI-generated speech can be exploited for malicious purposes

such as fraudulent phone calls to financial institutions, large-scale social engineering attacks, or automated voice scams that convincingly mimic trusted individuals [4], [5]. Fraud using AI-generated speech has led to multi-million dollar financial losses [6]. Furthermore, fraudulent calls and voice cloning can occur in real time during conversations with victims [7], [8], [9].

One effective solution to address these challenges is *watermarking* [10], [11], [12]. Audio watermarking systems embed imperceptible signals into audio content that can be identified by a corresponding detector mechanism. Specifically, modern audio watermarking systems alter the audio generation process so that all produced audio is automatically embedded with a unique footprint. Consequently, audio watermarks assist in detecting AI-generated audio, particularly in cases involving voice cloning, thereby preventing malicious actors from using voice cloning AI technology for fraud.

Research on audio watermarks advances in tandem with research on watermark removal. This close relationship is expected because without understanding how easily a watermark can be removed, users may develop a false sense of security and deploy ineffective watermarking solutions in real-world scenarios. Recent works on audio watermark removal have made important progress towards understanding the robustness of watermarking schemes via systematic evaluation of how resilient watermarks are to realistic signal perturbations and adversarial manipulations. However, existing approaches, ranging from basic signal-processing attacks to more advanced optimization-based methods, exhibit notable limitations. Signal-processing attacks, such as bandpass filtering or lossy compression [13], often fail to completely erase modern watermarks without introducing *audible artifacts*. Attack success rate is better for the optimization-based methods (e.g., HopSkipJump and Square attacks [14], [15]), but they lack practicality as they typically assume prior knowledge of the specific *watermark type* that each audio sample adopts or even *glass-box access* to the internal mechanisms of the watermark detector. Additionally, optimization-based methods require a large number of queries to the attacked watermarking scheme and are *costly in terms of attack time* [15]. Thus, these attacks fail to challenge the current defense mechanisms in realistic scenarios, particularly those requiring real-time watermark removal for voice cloning and impersonation [7], [8], [9].

To challenge state-of-the-art watermarks in realistic settings, this paper presents HarmonicAttack—the *first learning-based* audio watermark removal method, which inherently addresses the limitations that exist in current optimization-

<sup>†</sup>Equal contributions.

\*Under review.

based or signal-processing attacks. HarmonicAttack trains machine learning models to identify and disentangle watermark artifacts from the underlying audio content. In particular, we design a learning-based watermark removal method that (1) produces high-quality audio, (2) is not tailored exclusively to any specific watermarking scheme, (3) operates without glass-box access to the watermarking technique’s internals, (4) is computationally efficient and capable of real-time operations.

For practicality, we consider a closed-box threat model in which the adversary has access to the watermarking tool, but does *not* need access to internal parameters or the watermark detector. The attacker can embed watermarks into arbitrary audio samples and obtain a set of clean audio and corresponding watermarked audio pairs, drawn from some distribution, which need *not* match that of the target audio the attacker wishes to remove the watermarks from. This reflects a realistic deployment scenario where watermark embedding is publicly available (e.g., via an API), while the detection procedure remains private.

HarmonicAttack is an adversarially trained watermark removal approach that jointly trains a *watermark-removal generator* to learn watermark patterns from paired *clean* and *watermarked* audio data, and remove them, generating *unwatermarked* audio, and a *discriminator* (“adversarial” to generator) that distinguishes unwatermarked audio from authentic clean audio. The adversarial training, inspired by Generative Adversarial Networks (GANs) [16], is also referred to as GAN-style training [17], ensuring that generator outputs are statistically indistinguishable from authentic clean audio in the discriminator’s learned feature space, enhancing both watermark removal effectiveness and perceptual quality preservation achieved by the generator. To further improve watermark removal effectiveness and audio quality preservation capability, HarmonicAttack incorporates audio-specific architectural enhancements, including a *dual-path encoder* and *multiscale feature extraction* modules within the watermark-removal generator, as well as a novel *multi-component generator training loss* that integrates psychoacoustic principles to guide the watermark-removal generator towards perceptually consistent watermark removal.

Remarkably, we find that the exact choice of training dataset to which the watermark is applied is not critical for HarmonicAttack, as the watermark-removal generator transfers effectively to unseen audio datasets/domains and watermarking schemes. This robustness arises because watermark perturbations are largely constrained by perceptual masking rules [18], [19], based on the psychoacoustic principles. As a result, these perturbations tend to cluster within similar spectral regions, particularly around source-dominant frequency bands, allowing the removal model to generalize beyond its training distribution.

As a learning-based method, HarmonicAttack naturally requires time for training. However, this one-time effort enables it to achieve remarkable performance during inference. Unlike the state-of-the-art optimization-based attacks whose runtime grows with the input length [15], HarmonicAttack performs watermark removal in real time, with little sensi-

tivity to sample duration, achieving near-real-time removal even on long audio segments.

Together, these properties make HarmonicAttack not only a novel closed-box watermark removal approach but also an efficient and transferable solution for assessing the robustness of modern audio watermarking systems. To summarize, the paper makes the following contributions:

- We introduce the first learning-based audio watermark removal system that uses a single architecture to neutralize multiple state-of-the-art watermarking schemes (AudioSeal [10], WavMark [11], and Silentcipher [20]).
- We propose a neural network design combining a dual-path autoencoder, multiscale feature extraction, and GAN-style training to identify watermark patterns distributed across diverse time–frequency regions.
- We develop a novel multi-component loss that balances reconstruction fidelity with removal strength by emphasizing frequency bands where watermark energy is concentrated.
- We evaluate HarmonicAttack across audio domains and show strong transferrability, e.g., trained on speech, it achieves 100% watermark removal on unseen music from Free Music Archive (FMA) [21] with Perceptual Evaluation of Audio Quality (PEAQ) above 0.9, outperforming state-of-the-art removal methods.

We plan to make all artifacts, including the code, scripts, and datasets required to reproduce the results, publicly available.

## 2. Background

### 2.1. Audio Watermarking Schemes

The watermarking systems [10], [11] typically consist of an embedder that inserts the watermark and a detector that verifies its presence. Thus, when embedded in AI-generated audio, watermarks can help identify their origin.

The fundamental challenge in this domain lies in balancing two competing requirements: fidelity (maintaining audio quality so that the watermark is imperceptible), and robustness (surviving audio manipulations and performing consistently across different audio types, such as speech and music). Achieving higher robustness often means that the watermark has higher energy, which generally weakens its fidelity. Therefore, modern watermarking schemes aim to equalize both requirements by making subtle modifications to audio signals that take advantage of the limitations in human hearing. The human auditory system exhibits several psychoacoustic phenomena [19] that watermarking systems leverage, including frequency masking (where louder sounds mask quieter ones at nearby frequencies) and temporal masking (where sounds can mask other sounds that occur shortly before or after them).

Audio watermarking techniques can be categorized based on the domain in which they operate. Time-domain Watermarking directly modifies the audio waveform by adding watermark signals. Frequency-domain watermarking transforms the audio signal to reveal its frequency components, typically

using the Fast Fourier Transform (FFT) or Short-Time Fourier Transform (STFT) [22], and modifies the spectral coefficients. Transform-Domain Watermarking utilizes signal representations such as wavelet transforms, which can provide multi-resolution analysis. Modern watermarking methods are typically hybrid-domain. For instance, AudioSeal [10] uses a generator-detector architecture where the generator produces an additive watermark waveform directly in the time domain. However, its loss function accounts for the time-frequency features of the input, emphasizing the frequency-band signals where the original signal has high energy. This allows the approach to leverage the auditory masking property to improve fidelity.

## 2.2. Audio Watermark Removal

Approaches to watermark removal can be categorized based on their technical methodology. Conventional signal processing [13] employs operations like filtering, noise reduction, or codecs in an attempt to eliminate watermark components without prior knowledge of the watermarking system. Additive interference techniques [23], [24], [25] introduce noise patterns intended to disrupt the watermark detection without significantly degrading audio quality. Finally, optimization-based methods such as HopSkipJump and Square attacks, adapted to audio watermark removal in a recent study [15] iteratively optimize the noise patterns imposed on the watermarked samples. The optimization process is facilitated by the guidance from the watermark detection confidence score, accelerating the search process.

## 3. Design

HarmonicAttack uses an adversarial training setup inspired by GANs [16]: a *watermark-removal generator* that learns to remove watermarks from audio samples while a *discriminator* is trained simultaneously to distinguish *clean* audio from generator’s *unwatermarked* audio outputs.

The watermark-removal generator is a custom dual-path autoencoder that includes a waveform encoder, a spectrogram encoder, and a decoder. The watermark-removal generator takes both the clean and watermarked audio as input to learn the watermark patterns and remove them accordingly, producing unwatermarked audio.

Following the adversarial training paradigm, the discriminator is trained to perform the opposing task to the generator, namely, distinguishing unwatermarked audio produced by the generator from clean audio. Both components employ their respective training objectives during optimization. Specifically, for the watermark-removal generator, we design a novel *multi-component loss function* that balances watermark suppression and fidelity preservation, while the discriminator is trained using a standard binary cross-entropy classification loss. The system overview is shown in Figure 1.

### 3.1. Threat Model and Assumptions

We assume the adversary has access to: (1) a dataset of clean audio samples that importantly, need *not* to be of

same distribution as the watermarking scheme’s training set or the target test set, and (2) the ability to generate watermarked version of these samples using target watermarking systems (i.e., API access to the watermarking library without glass-box access to the full encoder/detector weights). The adversary’s goal is to remove watermarks from protected audio while maintaining perceptual quality.

### 3.2. Watermark-Removal Generator

Audio data can be represented in many domains, and the frequency and time domains are the most common domains for analysis. Audio in the time domain captures information like temporal patterns and amplitude variations over time, whereas frequency-domain audio representation reveals the spectral patterns of audio signals, which is complementary to the information provided in the time domain. Modern audio watermarks exploit both temporal and spectral characteristics of audio signals. AudioSeal [10], for instance, embeds watermarks across multiple frequency bands while maintaining temporal coherence. Single-domain processing (time-only or frequency-only) fails to capture the full scope of watermark embedding strategies. To address this, we propose a novel *dual-path autoencoder* architecture, modifying a standard convolutional autoencoder [26]. The dual-path autoencoder captures audio watermark information in both the time domain and frequency domain. As shown in Figure 2, the waveform encoder (top) extracts temporal features from the watermarked audio, while the spectrogram encoder (bottom) captures complementary spectral information from its STFT representation. Their outputs are concatenated in a shared bottleneck layer that learns the joint embeddings, which are then decoded by an attention-enabled decoder that produces unwatermarked audio.

**Waveform Encoder.** The *Waveform Encoder* (Figure 2, top blue trapezoid) processes 1D raw audio signals in the time domain through four convolutional layers [27]. Audio samples are standardized to a 16kHz sampling rate. Each convolutional layer incorporates 1D convolutions with decreasing kernel sizes to capture multiscale temporal patterns, along with batch normalization and residual connections to preserve fine-grained temporal information while maintaining training stability. This multiscale feature extraction pipeline captures watermark artifacts at different temporal scales (short-term frames for transient watermark patterns and long-term contexts for more persistent watermark residuals), utilizing the psychoacoustic principle that watermark residuals are embedded across various temporal resolutions to exploit auditory masking [18], [19]. This allows HarmonicAttack to generalize across different watermarking schemes that leverage different types of perceptual masking.

**Spectrogram Encoder.** The *Spectrogram Encoder* (Figure 2, bottom orange trapezoid) applies a Short-Time Fourier Transform (STFT) [22] to the watermarked audio and processes the resulting representation using 2D convolutional layers (as the audio signal becomes 2D data with both time and frequency dimensions). The STFT applied uses a 2048-point FFT with

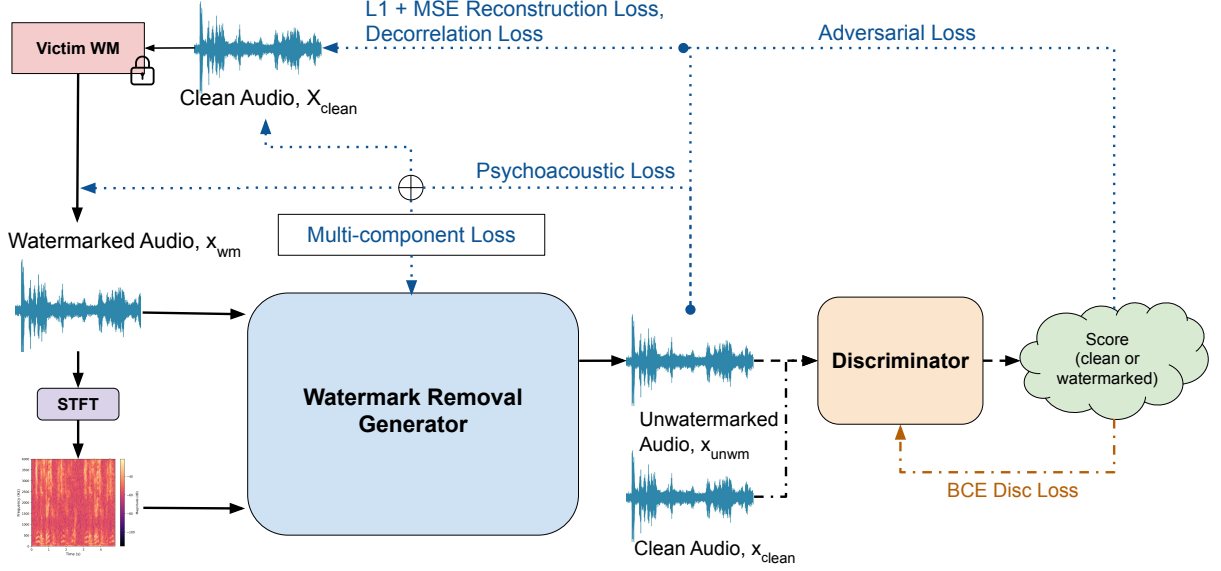


Figure 1. HarmonicAttack’s overview. The approach adopts a dual-path autoencoder architecture for the watermark-removal generator, and a discriminator for GAN-style adversarial training. The watermark-removal generator processes watermarked audio to produce unwatermarked outputs, while the discriminator learns to distinguish these from the corresponding clean references. The two models are co-trained iteratively, with the discriminator’s feedback guiding the generator towards improved watermark removal and perceptual fidelity.

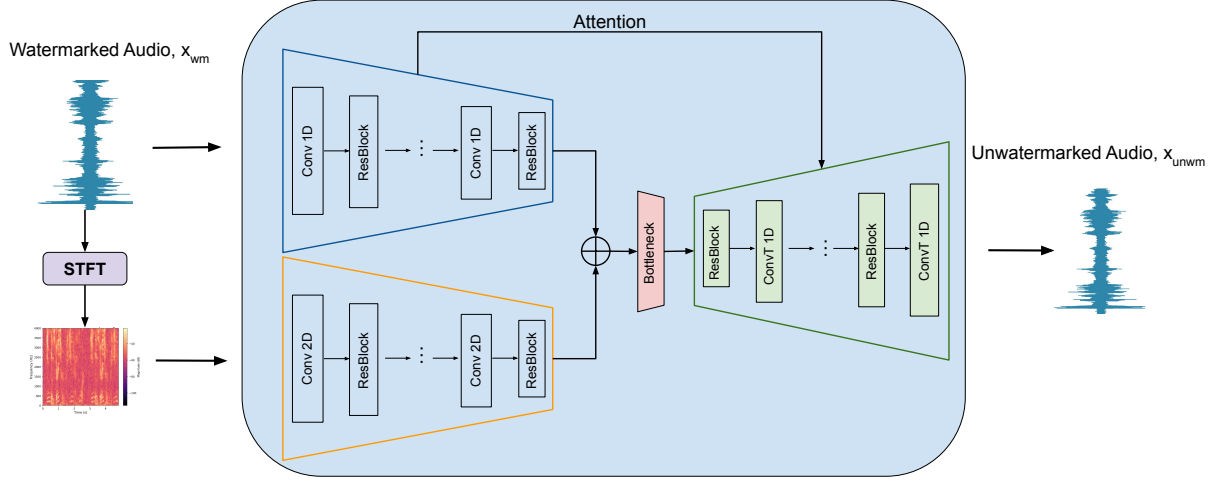


Figure 2. HarmonicAttack’s watermark-removal generator architecture.

a 512-sample hop length, generating spectrograms spanning 0-8kHz frequency range, which is a setting that balances resolution with efficiency. The spectrogram encoder utilizes the time–frequency domain of audio signals and captures watermark features that might not be detected through temporal analysis alone. Indeed, many audio denoising approaches also operate in the time–frequency domain [28], [29], which inspires our design choice. The architecture mirrors that of the waveform encoder but uses 2D operations suitable for time-frequency representations.

**Attention-Based Decoder.** In our decoder architecture (Figure 2, right green trapezoid), we enhance the standard convolutional autoencoder design by incorporating attention

mechanisms that establish selective connections between corresponding encoder and decoder layers, preserving critical long-range dependencies. Inspired by LightShed’s attention-based design [30], this architecture enables HarmonicAttack to focus computational resources on regions containing watermark artifacts while preserving the integrity of authentic audio content, thereby achieving more precise and targeted watermark removal.

### 3.3. Discriminator

Our discriminator is adversarial to the generator. The discriminator distinguishes between clean and unwatermarked

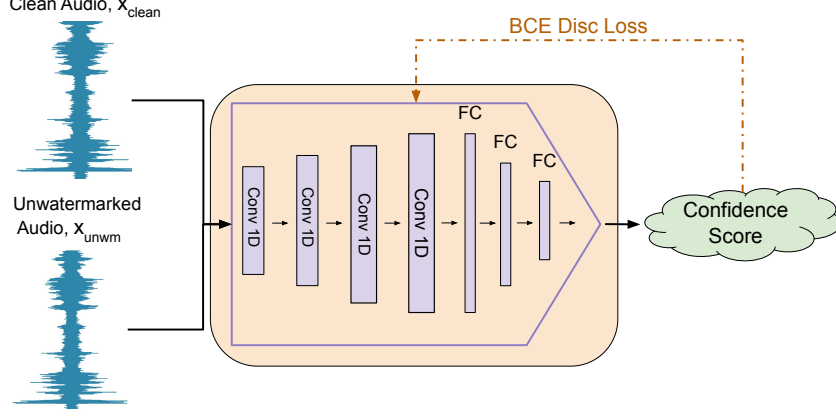


Figure 3. HarmonicAttack’s adversarial discriminator architecture.

audio (See Figure 3). It takes a single audio input and outputs a binary classification result. With the presence of the discriminator, the watermark-removal generator must produce unwatermarked outputs that confuse the discriminator into believing they are clean, i.e., never watermarked. Therefore the discriminator provides adversarial supervision that forces the watermark-removal generator to produce outputs indistinguishable from clean audio, preventing HarmonicAttack from applying aggressive noise or distortion that might remove watermarks but significantly degrade audio quality.

### 3.4. Loss Function Design

Our training objective for the watermark-removal generator employs a novel *multi-component loss* function designed to balance watermark removal effectiveness with perceptual quality preservation. Specifically, our training objective for the generator combines four losses in total, including reconstruction loss  $\mathcal{L}_{\text{recon}}$ , psychoacoustic loss  $\mathcal{L}_{\text{psychoacoustic}}$ , decorrelation loss  $\mathcal{L}_{\text{decorr}}$ , and adversarial loss  $\mathcal{L}_{\text{adv}}$ :

$$\mathcal{L}_{\text{total}} = \alpha_r \mathcal{L}_{\text{recon}} + \alpha_p \mathcal{L}_{\text{psychoacoustic}} + \alpha_{\text{wd}} \mathcal{L}_{\text{decorr}} + \alpha_a \mathcal{L}_{\text{adv}}$$

where  $\alpha_r$ ,  $\alpha_p$ ,  $\alpha_{\text{wd}}$ , and  $\alpha_a$  are the tunable hyperparameters that control the relative importance of reconstruction quality, psychoacoustic masking awareness, watermark removal strength and adversarial robustness, respectively.

**Generator Reconstruction Loss.** The reconstruction loss enforces temporal-domain fidelity by combining L1 and L2 distances between the unwatermarked audio (generator’s output) and the clean reference audio:

$$\mathcal{L}_{\text{recon}} = |x_{\text{unwm}} - x_{\text{clean}}| + 0.1(x_{\text{unwm}} - x_{\text{clean}})^2$$

The hybrid formulation leverages the robustness of L1 loss to outliers while incorporating the smoothness properties of L2 loss through a reduced weighting factor at 0.1.

This loss component provides the fundamental signal preservation constraint that prevents the model from introducing arbitrary distortions during watermark removal. Unlike frequency domain losses that focus on spectral

characteristics, reconstruction loss operates directly on the waveform samples, ensuring that the basic amplitude and phase relationships are maintained.

**Generator Psychoacoustic Loss.** The psychoacoustic loss exploits the fundamental principle that audio watermarks are strategically embedded in time-frequency regions where the original audio provides natural perceptual masking. Our approach inverts this masking strategy by learning and targeting these exact embedding locations:

$$\mathcal{L}_{\text{psychoacoustic}} = \sum_{m=1}^M w_m \cdot r_m$$

Where the attention weights are computed using a softmax normalization over watermark energy distributions:

$$w_m = \frac{\exp(e_m)}{\sum_{k=1}^M \exp(e_k)}$$

The watermark energy in mel band  $m$  is calculated as the temporal average of power spectral differences:

$$e_m = \frac{1}{T} \sum_{t=1}^T (\text{Mel}_m(\|\text{STFT}(x_{\text{wm}})\|^2)_t - \text{Mel}_m(\|\text{STFT}(x_{\text{clean}})\|^2)_t)$$

The processed residual energy represents the remaining artifacts after our removal process:

$$r_m = \frac{1}{T} \sum_{t=1}^T (\text{Mel}_m(\|\text{STFT}(x_{\text{unwm}})\|^2)_t - \text{Mel}_m(\|\text{STFT}(x_{\text{clean}})\|^2)_t)$$

Here,  $\text{Mel}(\cdot)$  represents the mel-scale frequency bands [31] covering the perceptually critical range of 200Hz-8kHz,  $m$  is an index of the mel frequency band (out of  $M$  total bands, each corresponding to a specific perceptual frequency range), and  $t$  represents discrete time frames in the STFT

analysis window. We split audio into mel bands because their separate processing provides two key advantages for watermark removal. First, mel bands mirror the non-linear frequency resolution of the human auditory system, where lower frequencies are processed with finer resolution than higher frequencies, naturally aligning our optimization with the perceptual significance of different spectral regions. Second, since audio watermarks exploit psychoacoustic masking [18], [19] by embedding in perceptually less sensitive regions, mel bands help identify the exact frequency ranges where watermarks are more likely embedded, which enables HarmonicAttack to achieve more precise removal.

The psychoacoustic loss component addresses the challenge that watermark removal is fundamentally different from general audio restoration. While reconstruction loss ensures global fidelity, it treats all time-frequency regions equally and cannot distinguish between watermark-carrying frequencies and natural audio content. Additionally, a uniform spectrogram reconstruction loss would waste optimization effort on frequencies that contain no watermark signatures, which can result in under-optimization of the key frequency regions where watermark signals are present. Gradients would be averaged across *all* frequencies, and watermark-relevant gradients would get diluted by non-watermark gradients. Our psychoacoustic loss fills this gap by providing an adaptive attention mechanism on important frequency bands, which focuses removal efforts specifically on perceptually masked regions where watermarks have the highest energy.

**Generator Decorrelation Loss.** The decorrelation loss ensures that the watermark-removal generator’s modifications are directly opposing the watermark patterns, enhancing the watermark removal strength:

$$\mathcal{L}_{\text{decorr}} = \frac{1}{2} (1 + \text{cosine\_sim}(\Delta_{\text{proc}}, \Delta_{\text{wm}}))$$

where  $\Delta_{\text{proc}} = x_{\text{unwm}} - x_{\text{clean}}$  and  $\Delta_{\text{wm}} = x_{\text{wm}} - x_{\text{clean}}$ . This loss component is responsible for addressing the limitations that reconstruction and psychoacoustic losses alone would pose. While preserving audio quality and targeting frequency-specific regions, they cannot guarantee that the watermark-removal generator’s learned transformations actually oppose watermark patterns. Therefore, this loss provides an alternative component in the gradient calculation that helps the training converge faster.

**Generator Adversarial Loss.** If watermarks are partially removed but the resulting audio exhibits statistical artifacts or unnatural characteristics detectable by machine learning systems, classifiers may still flag these samples as “unwatermarked” or “suspicious” instead of “clean” (never watermarked), even without detecting the embedded watermark. This motivates the inclusion of a discriminator network that enforces distributional realism, thereby enabling a GAN-style adversarial training paradigm.

The adversarial loss encourages the generator to produce audio that is statistically indistinguishable from natural, unprocessed audio in the learned feature space of the discriminator. This ensures that even when perfect reconstruction is not achieved by the watermark-removal generator, the

unwatermarked audio exhibits the same statistical properties and naturalness as authentic clean audio, making it sound more natural acoustically so that human perceptual system is less likely to identify processing evidences:

$$\mathcal{L}_{\text{adv}} = \text{BCE}(D(x_{\text{unwm}}), 1),$$

where  $D(\cdot)$  represents the discriminator output, BCE denotes binary cross-entropy loss that trains the watermark-removal generator to produce samples that fool the discriminator into classifying them as clean audio. The discriminator  $D$  is trained to distinguish clean audio from unwatermarked audio, which we discuss in detail in section 3.4. This setup creates a competitive training objective where the generator should produce outputs that are indistinguishable from natural, unprocessed audio.

The adversarial loss addresses the fundamental limitation that none of the other losses can overcome: ensuring that the unwatermarked audio sounds natural to human ears, even when perfect watermark residual removal is not achieved. The adversarial loss optimization objective operates through a learned discriminative signal that is given by the discriminator to the generator, making the optimization complementary to the constraints imposed by other loss components.

**Discriminator Training Loss.** The discriminator is trained together with the encoder using a standard binary cross-entropy loss that enforces correct classification between two categories, namely clean and unwatermarked audio (produced by the generator):

$$\mathcal{L}_{\text{disc}} = \frac{1}{2} [\mathcal{L}_{\text{clean}} + \mathcal{L}_{\text{unwatermarked}}],$$

where each component targets a specific audio category:

$$\begin{aligned} \mathcal{L}_{\text{clean}} &= \text{BCE}(D(x_{\text{clean}}), 1) \\ \mathcal{L}_{\text{unwatermarked}} &= \text{BCE}(D(G(x_{\text{wm}})), 0) \end{aligned}$$

Here,  $D(\cdot)$  represents the discriminator’s classification results,  $G(\cdot)$  denotes the watermark-removal generator, and BCE is the binary cross-entropy loss. The clean audio component trains the discriminator to recognize natural, unprocessed clean audio (label 1). The unwatermarked audio component ensures the discriminator can detect the watermark-removal generator’s generated outputs (label 0). This formulation creates a comprehensive training signal that enables the discriminator to distinguish clean audio from unwatermarked, providing robust adversarial pressure for the watermark-removal generator’s improvement.

### 3.5. Unified Training and Testing Loop for the Watermark-Removal Generator and Discriminator

We have two trainable components: (1) watermark-removal generator training with the custom multi-component loss—to establish watermark removal capabilities, and (2) adversarial training with corresponding discriminator training loss and the adversarial loss feeding back to the watermark-removal generator—to achieve robust watermark removal,

with respect to the adversarial discriminator. After all the models are trained, we test the attack effectiveness and audio quality on unseen test data. The training and testing algorithms are shown in Algorithms 1 and 2.

---

**Algorithm 1** HarmonicAttack Training

---

```

1: Input: Dataset  $\mathcal{D} = \{(x_i^{\text{clean}}, x_i^{\text{wm}})\}_{i=1}^N$ , epochs  $E$ ,
2: Output: Trained autoencoder  $G_\theta^*$ , discriminator  $D_\phi^*$ 
3: Initialize autoencoder  $G_\theta$ , discriminator  $D_\phi$ 
4: Initialize optimizers  $\text{opt}_G$ ,  $\text{opt}_D$ 
5: for epoch  $e = 1$  to  $E$  do
6:   for batch  $(x_{\text{clean}}, x_{\text{wm}})$  in  $\mathcal{D}$  do
7:      $\triangleright$  Train discriminator
8:      $x_{\text{unwm}} \leftarrow G_\theta(x_{\text{wm}})$   $\triangleright$  Generate unwatermarked
       audio
9:      $\mathcal{L}_{\text{clean}} \leftarrow \text{BCE}(D_\phi(x_{\text{clean}}), 1)$ 
10:     $\mathcal{L}_{\text{unwatermarked}} \leftarrow \text{BCE}(D_\phi(\text{detach}(x_{\text{unwm}})), 0)$ 
11:     $\mathcal{L}_{\text{disc}} \leftarrow \frac{1}{2}[\mathcal{L}_{\text{clean}} + \mathcal{L}_{\text{unwatermarked}}]$ 
12:    Update  $D_\phi$  with  $\nabla_{\phi} \mathcal{L}_{\text{disc}}$ 
13:     $\triangleright$  Train autoencoder
14:     $\mathcal{L}_{\text{recon}} \leftarrow |x_{\text{unwm}} - x_{\text{clean}}| + 0.1(x_{\text{unwm}} - x_{\text{clean}})^2$ 
15:     $\mathcal{L}_{\text{decorr}} = \frac{1}{2}(1 + \text{cosine\_sim}(\Delta_{\text{proc}}, \Delta_{\text{wm}}))$ 
16:     $\mathcal{L}_{\text{psychoacoustic}} \leftarrow \sum_{m=1}^M w_m \cdot r_m$ 
17:
18:     $\mathcal{L}_{\text{adv}} \leftarrow \text{BCE}(D_\phi(x_{\text{unwm}}), 1)$   $\triangleright$  Confuse
       discriminator
19:     $\mathcal{L}_{\text{total}} \leftarrow \alpha_r \mathcal{L}_{\text{recon}} + \alpha_p \mathcal{L}_{\text{psychoacoustic}} + \alpha_{\text{wd}} \mathcal{L}_{\text{decorr}} +$ 
        $\alpha_a \mathcal{L}_{\text{adv}}$ 
20:    Update  $G_\theta$  with  $\nabla_{\theta} \mathcal{L}_{\text{total}}$ 
21:  end for
22: end for
23: return  $G_\theta^*$ ,  $D_\phi^*$ 

```

---



---

**Algorithm 2** HarmonicAttack Testing

---

```

1: Input: Trained model  $G_\theta^*$ , watermarked audio  $x_{\text{wm}}$ ,
   victim detector  $V$ 
2: Output: Processed audio  $x_{\text{unwm}}$ , attack success flag
3:  $x_{\text{unwm}} \leftarrow G_\theta^*(x_{\text{wm}})$   $\triangleright$  Single forward pass
4:  $\text{conf}_{\text{before}} \leftarrow V.\text{detect}(x_{\text{wm}})$   $\triangleright$  Original detection
   confidence
5:  $\text{conf}_{\text{after}} \leftarrow V.\text{detect}(x_{\text{unwm}})$   $\triangleright$  Post-attack confidence
6:  $\text{success} \leftarrow (\text{conf}_{\text{after}} < \tau)$   $\triangleright$   $\tau$  is detection threshold
7: return  $x_{\text{unwm}}$ , success

```

---

## 4. Evaluation

In this section, we discuss the experimental evaluation of HarmonicAttack, focusing on its effectiveness across multiple victim watermarks and audio domains. We first compare its performance against baseline competitors on different victim watermark models and data types. We then examine the extent to which HarmonicAttack transfers to an unseen audio domain without retraining, highlighting its generalizability. Next, we analyze how audio sample length influences the performance of HarmonicAttack relative to existing methods.

Finally, we investigate the contribution of each architectural and loss component to understand which elements have the greatest impact on its watermark removal capability.

### 4.1. Experimental Setup

All experiments were performed on a system with two Intel Xeon 6548Y processors, 512GB of RAM and four Nvidia H100 GPUs having 96GB high-bandwidth memory.

**Datasets.** Our evaluation employs widely used audio datasets that represent diverse acoustic characteristics and usage scenarios commonly encountered in real-world watermarking applications. We utilize two primary datasets: LibriSpeech [32] for speech content and FMA-small, a subset of Free Music Archive (FMA) [21], for musical content, ensuring comprehensive coverage of different audio domains. We choose LibriSpeech due to its diverse speaker characteristics, standardized preprocessing, and consistent 16kHz sampling rate. The FMA-small dataset, which we refer to as the FMA dataset for brevity, in its turn, provides musical content with diverse genres, instruments, and acoustic characteristics. It contains 8,000 tracks across 8 genres, each 30 seconds in duration. Musical content poses unique challenges for watermark removal due to its complex spectral structure with multiple simultaneous instruments that creates rich masking opportunities for watermark embeddings. For our experiments, we use  $\mathcal{D}_{\text{train}}$  to denote the dataset we train the attack on and  $\mathcal{D}_{\text{eval}}$  to denote the dataset we evaluate the attack on.  $\mathcal{D}_{\text{train}}$  and  $\mathcal{D}_{\text{eval}}$  can be different datasets to study HarmonicAttack transferability (Section 4.3).

**Watermarking Methods.** We evaluate HarmonicAttack against three state-of-the-art audio watermarking systems, namely AudioSeal [10], WavMark [11], and Silentcipher [20]. AudioSeal, one of the most recent and robust audio watermarking methods, has demonstrated strong resilience against a comprehensive suite of signal-processing attacks. WavMark and Silentcipher are also contemporary and widely evaluated watermarking schemes, making them well-suited as representative watermarks for our evaluation. AudioSeal and Silentcipher are both neural audio watermarking systems that typically employ generator-detector architectures. However, the watermarks they embed have different capacities. Silentcipher has multi-bit capacity, embedding binary strings that can be decoded by the detector, whereas AudioSeal is a zero-bit scheme—its detector is a binary classifier that classifies a sample as watermarked when the confidence score is above 50%. In contrast, WavMark represents a traditional spread-spectrum watermarking approach embedding binary string watermarks directly into the frequency domain.

**Baseline Attacks.** We compare HarmonicAttack against two state-of-the-art attacks to demonstrate its effectiveness in attack success, resulting (unwatermarked) audio quality, and processing time. We picked as our primary baseline competitor the strongest-performing black-box attack from the recent paper assessing watermarking robustness [15], which we refer to as *AudioSquareAttack*. Based on the original image adversarial attack [14], AudioSquareAttack

uses optimization to find minimal perturbations that confuse audio watermark detectors. The attack employs a square-shaped perturbation pattern and iteratively refines the attack through querying the victim watermark detector.

Codec compression serves as our second competitor. It represents traditional signal processing attacks through lossy audio compression, as evaluated in recent work [13]. Following this work, we picked multiple codec configurations including traditional MP3 and OGG, and the neural EnCodec system developed by Meta [33] at various compression levels. Traditional codecs are tested at bitrates ranging from 32-320 kbps. For traditional codecs, we present the results of the codec/bitrate pair that yields the highest ASR in the experimental results. EnCodec is evaluated at multiple bandwidth settings, and the results are reported for the best-performing configuration. All codecs process audio through a compression-decompression cycle.

**Evaluation Metrics.** To evaluate HarmonicAttack and the baselines, we utilize metrics that measure the watermark removal effectiveness and efficiency, and the audio quality preservation capability. The metrics we adopt are as follows:

- 1) **Attack Success Rate (ASR) ( $\uparrow$ ).** For AudioSeal, we consider a watermark removal attempt successful if the watermark detector’s confidence score drops below the detection threshold (i.e., 0.5) after processing. For WavMark and Silencipher, the attack is successful when the decoded message is incomplete or different from the original embedded message. ASR is calculated as the proportion of watermarked audio samples where the watermark becomes undetectable after applying a removal method.
- 2) **Perceptual Evaluation of Audio Quality (PEAQ) ( $\uparrow$ ).** PEAQ is a standard metric that provides a perceptual assessment of audio quality, aiming to correlate with human auditory perception. PEAQ values range from 0 to 1, where values closer to 1 indicate better preserved perceptual quality. This metric is important as it reflects the actual listening experience quality rather than purely technical signal measurements.
- 3) **Attack Time ( $\downarrow$ ).** Attack time metric measures the average time in seconds required to complete the watermark removal process for a single audio sample, reflecting the computational efficiency and practicality of the attack method in real-time scenarios.

## 4.2. Performance Evaluation

We first use the LibriSpeech dataset to train HarmonicAttack and evaluate HarmonicAttack’s performance against the baselines on unseen samples in LibriSpeech. Table 1 shows the results of this evaluation. We then conduct comprehensive cross-dataset transfer experiments, evaluating HarmonicAttack’s generalizability. That is, we train HarmonicAttack’s models on one dataset and evaluate them on the other, demonstrating HarmonicAttack’s ability to learn domain-agnostic watermark removal strategies. Specifically, we train on LibriSpeech speech data and test transfer performance

TABLE 1. PERFORMANCE OF HARMONICATTACK AGAINST EXISTING COMPETITORS ON THE LIBRISPEECH DATASET. HARMONICATTACK IS TRAINED AND EVALUATED ON LIBRISPEECH DATASET, THE SAME AS THE BASELINE ATTACK COMPETITORS.

Attack	WM	ASR (%)	PEAQ	ATK Time (s)
MP3/OGG	AudioSeal	16	0.944	0.286
	WavMark	42	0.971	0.302
	Silencipher	100	0.966	0.186
EnCodec	AudioSeal	16	0.969	0.219
	WavMark	100	0.921	0.217
	Silencipher	100	0.925	0.251
AudioSquareAttack	AudioSeal	100	0.829	16.112
	WavMark	100	0.934	4.178
	Silencipher	100	0.886	13.64
Ours	AudioSeal	100	0.896	0.035
	WavMark	95	0.953	0.041
	Silencipher	100	0.915	0.059

TABLE 2. PERFORMANCE OF HARMONICATTACK AGAINST EXISTING BASELINE COMPETITORS ON FMA DATASET ( $\mathcal{D}_{\text{EVAL}}$ ). TO DEMONSTRATE TRANSFER, HARMONICATTACK IS TRAINED ON LIBRISPEECH DATASET ( $\mathcal{D}_{\text{TRAIN}}$ ) AND EVALUATED ON  $\mathcal{D}_{\text{EVAL}}$ .

Attack	WM	ASR (%)	PEAQ	ATK Time (s)
MP3/OGG (No transfer)	AudioSeal	0	0.926	0.241
	WavMark	48	0.960	0.182
	Silencipher	88	0.932	0.274
EnCodec (No transfer)	AudioSeal	0	0.928	0.550
	WavMark	100	0.820	0.524
	Silencipher	100	0.929	0.535
AudioSquareAttack (No transfer)	AudioSeal	2	0.767	102.275
	WavMark	44	0.890	34.039
	Silencipher	100	0.886	13.64
HarmonicAttack (LibriSpeech->FMA)	AudioSeal	100	0.900	0.029
	WavMark	78	0.901	0.058
	Silencipher	100	0.906	0.062

on FMA music samples, and vice versa, to assess HarmonicAttack’s ability to remove watermarks across audio with fundamentally different characteristics. The results of this evaluation are presented in Tables 2 and 3.

**HarmonicAttack.** HarmonicAttack demonstrates consistent performance across both audio domains. As shown in Table 1 and Table 2, HarmonicAttack achieves 100% ASR on both LibriSpeech and FMA against AudioSeal while maintaining superior PEAQ scores and significantly reduced attack times compared to AudioSquareAttack. Notably, the attack time remains consistent with that observed for speech sample transfers in Table 3 and is even five to ten faster than codec baselines, demonstrating that HarmonicAttack achieves real-time inference regardless of audio length or spectral complexity. This uniform computational efficiency stems from our model’s learned representations that generalize



TABLE 3. PERFORMANCE OF HARMONICATTACK AGAINST EXISTING BASELINE COMPETITORS ON LIBRISPEECH DATASET ( $\mathcal{D}_{\text{EVAL}}$ ). TO DEMONSTRATE TRANSFER, HARMONICATTACK IS TRAINED ON THE FMA DATASET ( $\mathcal{D}_{\text{TRAIN}}$ ) AND EVALUATED ON  $\mathcal{D}_{\text{EVAL}}$ .

Attack	WM	ASR (%)	PEAQ	ATK Time (s)
MP3/OGG (No transfer)	AudioSeal	16	0.944	0.286
	WavMark	42	0.971	0.302
	Silencipher	100	0.966	0.186
EnCodec (No transfer)	AudioSeal	16	0.969	0.219
	WavMark	100	0.921	0.217
	Silencipher	100	0.925	0.251
AudioSquareAttack (No transfer)	AudioSeal	100	0.829	16.112
	WavMark	100	0.934	4.178
	Silencipher	100	0.840	5.115
HarmonicAttack (FMA->LibriSpeech)	AudioSeal	100	0.802	0.0318
	WavMark	100	0.831	0.0330
	Silencipher	100	0.803	0.0180

across audio domains, eliminating the need for per-sample optimization inherent in baseline approaches.

**AudioSquareAttack.** In contrast, AudioSquareAttack faces significant limitations in this multi-domain scenario due to its non-adaptive, per-sample optimization nature. When evaluated on LibriSpeech speech samples, AudioSquareAttack achieves reasonable performance, demonstrating strong ASR on all three victims, only the attack time for each sample stays exceptionally high, around 4-16s, which violates the real-time requirement for attack scenarios like fraudulent calls and voice cloning, where real-time watermark removal is required. However, AudioSquareAttack’s performance significantly degrades when evaluated on the FMA music dataset. As shown in Table 2, AudioSquareAttack only achieves 2% ASR on the FMA dataset and AudioSeal, indicating that it is ineffective on music data. Additionally, it takes the attack around five to ten times longer to optimize on music samples compared to speech samples under the same number of attack iterations. This result shows the inconsistent performance of per-sample closed-box attack when facing diverse audio characteristics across speech and music domains.

Analysis of AudioSquareAttack’s evaluation results leads to the following conclusions. First, impractically, the attack cannot effectively operate without knowing in advance which method was used to watermark an audio sample. Second, domain-specific performance drop occurs for AudioSquareAttack because this method lacks knowledge about watermark patterns. AudioSquareAttack relies on iteratively adding random square noises to the watermarked audio and, in each step, querying the watermark detector to determine whether its detection confidence has dropped. This approach is highly sensitive to the audio domain characteristics. While square wave perturbations may effectively disrupt watermark detection in the speech audio data, which has relatively stable and simple spectral content, they prove inadequate for music samples with rich harmonic structures. Third, computation inefficiency becomes prohibitive as each audio

sample requires hundreds or thousands of optimization iterations, all requiring watermark detector queries, leading to attack times that scale linearly with audio length and dataset  $\mathcal{D}_{\text{eval}}$  size.

**Codecs.** Codec-based attacks also exhibit inconsistent performance that varies based on both the victim watermarking method and audio domain characteristics. Traditional codecs (MP3/OGG compression) demonstrate almost complete failure against robust watermarks like AudioSeal, achieving 0% and 16% ASR across music and speech datasets, indicating that modern neural watermarking methods are specifically designed to withstand traditional lossy compression.

The newer EnCodec neural codec shows improved removal performance, achieving high ASR against both WavMark and Silencipher, but still fails against AudioSeal’s robust embedding strategy. The higher ASR of EnCodec on WavMark in Table 2 is due to WavMark’s weaker robustness and EnCodec’s stronger compression, which removes watermarks but lowers PEAQ. This victim-specific performance disparity demonstrates a fundamental limitation of codec-based approaches: they rely on generic compression artifacts that may coincidentally disrupt some watermarking schemes while being ineffective on others that incorporate compression robustness into training objectives.

In contrast, HarmonicAttack shows consistently outstanding performance on all victim methods. Our attack achieves a lower ASR on WavMark when evaluated on FMA because, although some watermarking methods often embed signals in overlapping time–frequency regions to exploit the auditory masking effect, their specific embedding distributions still differ. Consequently, watermark removal becomes imperfect for certain samples when the distributions differ sufficiently. This transfer result on WavMark involves transferring both the dataset (from LibriSpeech to FMA) and the victim watermarking method (from AudioSeal to WavMark), leading to a larger distribution shift. When HarmonicAttack is trained on watermarked audio samples from the same distribution as the audio samples to be attacked, as shown in Table 1, ASR is comparable to EnCodec with higher PEAQ, even though the watermarking scheme being attacked may be different from the one being trained on.

### 4.3. Transferability Analysis

In this section, we analyze HarmonicAttack’s and AudioSquareAttack’s behavior during the evaluation to identify the reasons why HarmonicAttack demonstrates strong transferability across various audio data types. In particular, HarmonicAttack’s removal model trained only on speech data in LibriSpeech transfers to music audio data in FMA and vice versa. In contrast, although AudioSquareAttack does not need to train a model using specific data, its performance is largely dependent on the target audio data type.

We attribute this strong generalization capability to the ability of HarmonicAttack to locate regions in the time-frequency space of the target audio sample, where the watermark residual energy is high, due to several architectural

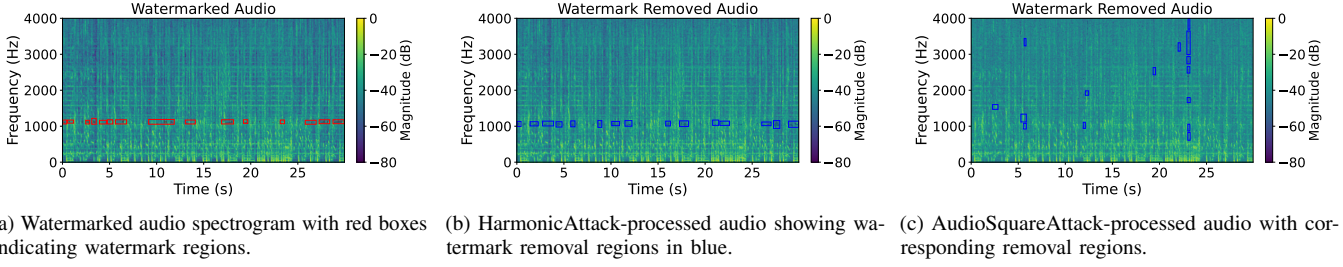


Figure 4. Comparison of spectrograms for watermarked audio, HarmonicAttack removal, and AudioSquareAttack removal on FMA AudioSeal sample. HarmonicAttack is evaluated by transferring from the model trained on AudioSeal LibriSpeech samples.

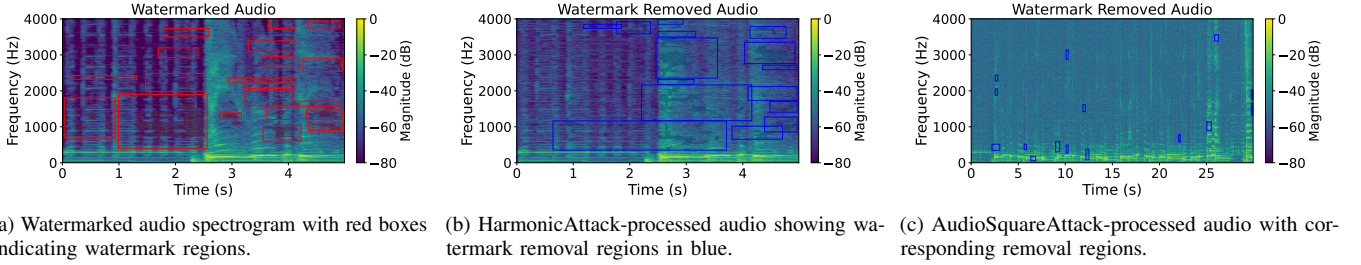


Figure 5. Comparison of spectrograms for watermarked audio, HarmonicAttack removal, and AudioSquareAttack removal on FMA WavMark sample. HarmonicAttack is evaluated by transferring from the model trained on AudioSeal LibriSpeech samples.

advantages. First of all, our multiscale resolution encoder captures both fine-grained temporal patterns (crucial for speech) and coarse-grained structures (crucial for music). This multiscale approach enables the model to identify watermark signatures across different audio modalities without requiring domain-specific retraining. Second, the attention-augmented decoder dynamically focuses on spectrally anomalous regions regardless of audio content type. Unlike AudioSquareAttack’s fixed adversarial perturbations, our attention mechanism learns to identify watermark-induced spectral deviations that manifest consistently across different audio domains. Third, our psychoacoustic loss, with frequency-adaptive weighting, exploits the fact that many watermarking schemes embed signatures in perceptually less sensitive frequency regions or regions where the acoustic masking effect [18], [19] is strong. This phenomenon remains consistent across audio types, where HarmonicAttack internally learns to identify these regions based on the training data.

To confirm our intuitions, we perform spectrogram analysis. We first select the original FMA and the corresponding AudioSeal-watermarked samples and locate where the watermarks live. Then, to identify whether HarmonicAttack and AudioSquareAttack target these watermarked regions, we compare the spectrograms of watermarked and the corresponding unwatermarked audio. Figures 4a, 4b and 4c show time-frequency spectrograms generated using STFT analysis with normalized magnitude values displayed in dB scale (0 to -80 dB range), with highlighted regions indicating watermark locations and removal patterns overlaid on the spectral representations. These spectrograms demonstrate that HarmonicAttack successfully identifies the watermark

regions and remove the watermark signals in those regions, while AudioSquareAttack remove signals spread across the whole time-frequency space and fails to focus on the watermark regions. This explains why AudioSquareAttack performs well on speech samples but fails on music samples: the search space for music samples is much larger than speech samples as music samples contain multiple instruments and voice tracks, each operating in a different frequency band. Therefore, AudioSquareAttack’s adversarial search becomes exponentially more challenging in the complex spectral landscape of music with overlapping instrumental harmonics and concurrent melodic lines. For completeness, we further provide plots that show the spectrograms of watermark signal before and after removal by HarmonicAttack and AudioSquareAttack in Figure 6 in Appendix A.

We also analyze the spectrograms of WavMark-watermarked audio samples presented in Figures 5a, 5b, and 5c. The corresponding spectrograms of watermark signals are in Figure 7 in Appendix A. These spectrograms provide two key observations supporting the transferability discussion of HarmonicAttack, across different datasets and different victim watermarking schemes. First, although watermarking schemes use distinct embedding heuristics, they are all constrained by and exploit the auditory masking effect [18], [19]. As a result, watermark residuals tend to concentrate in similar frequency regions (typically around 1000 Hz). This common structure enables HarmonicAttack, trained solely on AudioSeal, to generalize to unseen schemes such as WavMark. Second, WavMark’s watermark residuals are spread much more broadly across the time-frequency domain, whereas AudioSeal produces more localized residuals. This

TABLE 4. IMPACT OF FMA SAMPLE LENGTH ON HARMONICATTACK’S EFFECTIVENESS COMPARED TO AUDIOSQUAREATTACK. HARMONICATTACK IS TRAINED ON LIBRISPEECH AND EVALUATED ON FMA.

Length (s)	Method	ASR (%)	PEAQ	ATK Time (s)
1	AudioSquareAttack	38	0.764	18.07
	HarmonicAttack	98	0.899	0.0196
5	AudioSquareAttack	12	0.764	44.39
	HarmonicAttack	100	0.918	0.0354
10	AudioSquareAttack	8	0.759	69.01
	HarmonicAttack	100	0.909	0.0367
15	AudioSquareAttack	6	0.758	78.48
	HarmonicAttack	100	0.907	0.0308
30	AudioSquareAttack	2	0.767	102.275
	HarmonicAttack	100	0.900	0.0299

phenomenon arises because AudioSeal employs a time-frequency loudness loss (TF-Loudness) that concentrates watermark energy in narrow, perceptually optimal locations. In contrast, WavMark utilizes invertible neural networks that operate across the entire spectrogram, spreading watermark information more broadly across frequency and time dimensions. Although AudioSquareAttack is consistently weaker than HarmonicAttack on both AudioSeal and WavMark when evaluated on FMA, its use of random square-noise perturbations increases the chance of overlapping WavMark’s more widely distributed watermark patterns. Consequently, AudioSquareAttack performs better on WavMark than on AudioSeal, as shown in Table 2.

#### 4.4. Impact of Audio Sample Length

The results in Table 4 demonstrate how audio sample length affects the performance of HarmonicAttack versus AudioSquareAttack when attacking AudioSeal watermarks on FMA music samples. Notably, AudioSquareAttack’s cost increases dramatically with the length of the audio sample while HarmonicAttack’s cost increases only slightly.

HarmonicAttack achieves high and stable ASR (98-100%) while preserving excellent audio quality (PEAQ scores above 0.9) and maintaining real-time fast inference (around 0.03 seconds) across all tested audio lengths. In contrast, AudioSquareAttack exhibits severe performance degradation as sample length increases, with ASR dropping from 38% on 1-second clips to merely 2% on 30-second samples. Simultaneously, AudioSquareAttack’s computational cost increases dramatically with length, requiring a 466% increase in attack time: from 18 seconds for 1-second audio to over 102 seconds for 30-second samples.

This length-dependent degradation of AudioSquareAttack highlights a critical limitation for its real-world deployment, where audio content typically exceeds 30 seconds. Music tracks usually range from 3 to 5 minutes, podcasts and lectures can span hours, and streaming applications process continuous audio streams. Our results suggest that AudioSquareAttack’s effectiveness would further degrade

on much longer content, potentially requiring prohibitively long attack times while achieving minimal success rates. This phenomenon stems from the nature of the closed-box operation AudioSquareAttack employs, which applies random square-wave perturbations across the entire audio spectrogram in an attempt to disrupt watermark-embedded regions. As audio length and spectral complexity increase, the search space expands exponentially, making it increasingly difficult for the random perturbations to effectively target the specific frequency-time locations where watermarks are embedded. Furthermore, due to the high computational overhead of AudioSquareAttack, it cannot be used in real-time voice cloning scenarios, which limits its ability to challenge or validate current watermarking defenses.

In contrast, HarmonicAttack’s consistent performance across all tested sample lengths, combined with its real-time inference regardless of the duration, positions it as a more practical solution for real-world watermark removal scenarios.

TABLE 5. ABLATION STUDY OF HARMONICATTACK’S ARCHITECTURAL COMPONENTS EVALUATED ON LIBRISPEECH, WHERE  $\mathcal{D}_{\text{TRAIN}}$  IS LIBRISPEECH AND VICTIM IS AUDIOSEAL.

Ablation	ASR (%)	PEAQ	ATK Time (s)
Baseline	71	0.882	0.036
Adv	100	0.896	0.035
Baseline-D	93	0.884	0.034
Adv-D	93	0.920	0.037

#### 4.5. Ablation Study

**Effect of Each Architectural Component.** In order to study the effect of individual components in HarmonicAttack’s architecture (Figure 1), we conduct an ablation study. Specifically, we evaluate our design in the following settings:

- 1) *Baseline*: Watermark-Removal Generator only
- 2) *Adversarial (Adv)*: Watermark-Removal Generator and Discriminator—this is our main setting.
- 3) *Baseline with detector support (Baseline-D)*: Detector-confidence-score-guided Watermark-Removal Generator
- 4) *Adversarial with detector support (Adv-D)*: Detector-confidence-score-guided Watermark-Removal Generator and Discriminator

For this set of experiments, we set the training dataset  $\mathcal{D}_{\text{train}}$  and the testing dataset  $\mathcal{D}_{\text{eval}}$  to LibriSpeech. The audio is watermarked with AudioSeal.

The results are presented in Table 5. As expected, incorporating the discriminator (*Adv*) component increases the ASR from 71% to 100%, demonstrating that the discriminator-guided generator can remove embedded watermarks more effectively. At the same time, it preserves perceptual quality (PEAQ is 0.896 versus the baseline 0.882), with negligible impact on attack time. Since the AudioSquareAttack assumes access to detectors, we also test the effectiveness of HarmonicAttack having the closed-box detector access (i.e., confidence score only). Adding detector guidance

(*Baseline-D*) also boosts ASR relative to the baseline (93% vs. 71%), indicating that the detector provides informative feedback that helps the generator find watermark-related regions more effectively. Interestingly, when combining both discriminator and detector (*Adv-D*), while ASR remains 93%, its perceptual quality improves slightly (i.e., PEAQ increased to 0.920). This suggests that the detector confidence scores introduce a regularizing effect when collaborating with the discriminator. Overall, the results confirm that both discriminator and detector supervisions are complementary to generator in enhancing HarmonicAttack’s attack performance. The discriminator primarily improves watermark removal, while the detector feedback stabilizes training and refines perceptual quality.

TABLE 6. ABLATION STUDY OF HARMONICATTACK (*Adv*)’S MULTI-OBJECTIVE LOSS FUNCTION EVALUATED ON LIBRISPEECH, WHERE  $\mathcal{D}_{\text{TRAIN}}$  IS LIBRISPEECH AND VICTIM IS AUDIOSEAL.

Ablation	ASR (%)	PEAQ	ATK Time (s)
$\mathcal{L}_{\text{recon}}$ excluded	99	0.791	0.035
$\mathcal{L}_{\text{psychoacoustic}}$ excluded	84	0.963	0.036
$\mathcal{L}_{\text{decorr}}$ excluded	66	0.848	0.034
All	100	0.896	0.035

**Effect of Each Loss Component.** To understand the contribution of each term in HarmonicAttack’s multi-objective loss function, we perform an ablation study by selectively removing individual loss components while keeping all other training settings fixed. The experiments are conducted on AudioSeal with LibriSpeech as  $\mathcal{D}_{\text{train}}$  and  $\mathcal{D}_{\text{eval}}$ .

Table 6 summarizes the results. They demonstrate that removing  $\mathcal{L}_{\text{recon}}$  loss component slightly reduces the ASR from 100% to 99% and causes a notable drop in perceptual quality (PEAQ drops from 0.896 to 0.791), indicating that reconstruction loss is important for maintaining audio fidelity. The exclusion of  $\mathcal{L}_{\text{psychoacoustic}}$  loss, which strategically helps to locate watermarks in frequency-time regions following natural perceptual masking rules of watermarks [18], [19], decreases ASR to 84% while increasing PEAQ to 0.963. As expected, HarmonicAttack is more aggressive in watermark removal when psychoacoustic loss is used. Dropping  $\mathcal{L}_{\text{decorr}}$  loss component significantly degrades both ASR (66%) and PEAQ (0.848), showing that the decorrelation loss plays a key role in disentangling watermark-relevant and watermark-irrelevant representations within the watermark-removal generator. This loss encourages the attack’s perturbations to be decorrelated from the watermark embedding patterns, ensuring that HarmonicAttack does not align with or reinforce watermark features. Without it, the processed perturbations may partially overlap with the watermark subspace, which weakens watermark removal. Overall, the results demonstrate that those three loss components are complementary:  $\mathcal{L}_{\text{recon}}$  preserves perceptual quality,  $\mathcal{L}_{\text{psychoacoustic}}$  helps locate watermarks in the frequency bands, and  $\mathcal{L}_{\text{decorr}}$  enhances watermark feature isolation. Their joint optimization enables HarmonicAttack to achieve high ASR while producing high-fidelity audio.

Details on the hyperparameter search adjusting the significance of each loss components are discussed in Appendix B.

## 5. Discussion

### 5.1. Defenses against HarmonicAttack

HarmonicAttack is the first learning-based attack demonstrating that modern audio watermarks are susceptible to removal in realistic closed-box scenarios. Thus, there is a need for research to design effective watermarks capable of counteracting such removal.

From our empirical analyses of spectrograms before and after watermark removal, we observe distinct regions where schemes inject watermarks and where our attack successfully suppresses watermark-related signal perturbations. These findings indicate that the spectrogram representation preserves rich structural and statistical information, offering a sufficiently expressive space for watermark embedding. Building upon this insight, a promising line of defense is to leverage the time–frequency domain by transforming audio signals into spectrograms and adapting the ideas derived from advanced image watermarking schemes to this domain. Image watermarking is more explored than audio watermarking, and includes latent-space and semantic watermarking schemes that achieve high robustness against diverse manipulations [34], [35], [36]. Such cross-modal adaptation has the potential to integrate the robustness of image watermarking into the audio modality.

### 5.2. Watermarked Audio Sample Length

To stay consistent with the evaluation of the baseline watermarking schemes, which are typically evaluated on the LibriSpeech dataset using 5-second audio segments, we also restricted our evaluation of HarmonicAttack to 5-second LibriSpeech samples. While these samples are relatively short, our framework can be readily extended to longer speech datasets, such as [37]. Based on the results reported in Table 4, we anticipate that HarmonicAttack will perform effectively on longer audio segments. In contrast, AudioSquareAttack is likely to experience a significant decline in performance on extended speech samples, as suggested by the trends observed in Table 4. In real-world scenarios, audio samples are typically much longer. For example, music tracks often span 3–5 minutes, while speech recordings including lectures, conversations, or interviews can extend significantly beyond that.

### 5.3. Limitations

HarmonicAttack demonstrates strong transferability across a range of state-of-the-art watermarking schemes due to shared algorithmic principles, i.e., watermarks’ perturbations follow perceptual masking rules [18], [19]. As observed in the spectrogram analysis in Section 4.3, a model trained on AudioSeal effectively transfers to WavMark because

the majority of WavMark’s watermark-embedded regions spatially overlap with those of AudioSeal, allowing HarmonicAttack to exploit similar spectral and temporal patterns for watermark removal. However, such generalization is not guaranteed for all possible watermarking schemes and data samples, especially in scenarios where degrading the protected audio quality is acceptable. In such cases, additional training or finetuning on samples from the new scheme would be required to adapt the model effectively.

## 6. Related Work

### 6.1. Audio Watermarking

Recent advances in generative audio models have made synthetic content nearly indistinguishable from human-created works, raising growing concerns about provenance, copyright, and authenticity. To address these issues, researchers have developed watermarking techniques that embed imperceptible but detectable signals into generated media. These watermarking schemes can be categorized into multi-bit watermarks, which embed recoverable binary messages that the detector decodes, and zero-bit watermarks, in which the detector operates as a binary classifier that determines whether a watermark is present in the audio.

Among the zero-bit watermarks, there are AudioSeal [10], which represents a major step toward localized watermarking. It jointly trains a generator–detector pair where the generator produces an additive watermark waveform, and the detector estimates probabilities of watermark presence in each localized sample segment. The model is optimized under a combination of perceptual and detection losses. This setup yields watermarks that are both imperceptible and robust to common audio transformations. Building upon AudioSeal, Latent Watermarking of Audio Generative Models [38] extends watermarking from the waveform level to the latent space of audio language models. Unlike post-hoc watermarking, this method embeds watermarks into the training data of the audio model itself, ensuring that any model trained on such data inherently produces watermarked outputs. Audio Watermark [39] introduces a dynamic, style-transfer-based watermarking framework designed for black-box voice dataset ownership verification. By leveraging out-of-domain features and bi-level adversarial optimization, it produces harmless watermarks that preserve the original label while maintaining high verification accuracy.

Among the multi-bit watermarks, Invertible Audio Watermarking [11] reframes watermarking as a reversible transformation task using invertible neural networks (INNs). It encodes binary messages into spectrograms via STFT and jointly processes audio and watermark features through INN blocks, ensuring message recovery. Timbre Watermarking [12] embeds watermark information directly into the frequency domain of a speaker’s voice, leveraging repeated embedding to enhance robustness against common audio preprocessing. GROOT [40] introduces a generative audio watermarking paradigm that embeds watermarks directly

within diffusion-based vocoders by injecting a learned latent watermark vector into the model’s input noise space. By jointly training an encoder–decoder around a frozen diffusion model, GROOT enables plug-and-play watermark synthesis.

### 6.2. Audio Watermark Removal

Watermarking schemes are susceptible to removal attacks that can erase the embedded patterns. The existing attacks can be categorized into signal-processing-based attacks and optimization-based attacks based on the methodology.

Signal-processing-based attacks involves conventional signal processing techniques, such as bandpass filters or lossy compression codecs, to process watermarked samples aiming to distort the watermark patterns embedded in the audio. Prior studies [13], [41] have evaluated the effectiveness of these transformations against various watermarking schemes. However, such attacks often fail to fully remove modern, robust watermarks without introducing audible artifacts.

Optimization-based attacks form a more advanced class of watermark removal methods in which the attack perturbations are explicitly optimized using feedback from the watermark detector’s confidence scores. In contrast to signal-processing attacks that apply fixed transformations, these approaches craft a sample-specific perturbation tailored to each watermarked input, aiming for stronger removal and better perceptual quality. [15] survey a range of such methods, from closed-box square attacks [14] to glass-box adversarial attacks that rely on full knowledge of the detector’s parameters. These glass-box techniques assume prior knowledge of the watermarking scheme, such as the watermark type for each sample or access to all hyperparameters. For fairness, we compare against the strongest closed-box method presented in the survey—the square attack—which stays within our threat model, although it still operates under stronger assumptions by requiring repeated queries to the watermark detector.

## 7. Conclusion

In this work, we introduce HarmonicAttack, the first learning-based audio watermark removal attack. Unlike prior attacks that rely on glass-box access, extensive querying, and high computational overhead, our attack operates under realistic adversarial scenarios, e.g., real-time impersonalization/voice cloning, while preserving high audio quality. HarmonicAttack is closed-box and generalizable. It offers an effective, efficient, and cross-domain solution to evaluate modern audio watermarking schemes. Empirical results show that HarmonicAttack neutralizes state-of-the-art watermarking schemes, achieving mostly 100% ASR, near-real-time performance, and high quality of unwatermarked audio across both speech and music samples, even on long audio segments. In contrast, existing attacks often fail to remove watermarks, show reduced effectiveness on longer audio, or require excessively long runtimes as segment duration increases. These results demonstrate that current watermarking schemes, though robust against existing traditional and

optimization-based attacks, remain vulnerable to our learning-based HarmonicAttack. In summary, HarmonicAttack reveals structural weaknesses in modern watermarking methods and emphasizes the need for future designs that incorporate adaptive, adversarially robust defenses to ensure reliable audio watermarking in the era of generative media.

## Ethics Considerations

Advances in audio watermark removal are increasing the risks. For instance, they allow for the erasure of provenance from the AI-generated speech, facilitating misinformation, impersonation, and fraud. Still, researching watermark removal is ethically justified and important. By systematically studying the vulnerabilities of current watermarking schemes, we can identify weaknesses that might otherwise be exploited maliciously. Moreover, understanding these weaknesses enables the development of more robust watermarking techniques. Furthermore, publishing responsible findings encourages model owners to adopt stronger watermarking standards and promotes user awareness of the limitations of the existing protection mechanisms. Ultimately, such research contributes to a safer and more trustworthy digital ecosystem rather than undermining it.

## LLM Usage Considerations

LLMs were used for editorial purposes in this manuscript, and all outputs were inspected by the authors to ensure accuracy and originality.

## References

- [1] KimiTeam, D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, Z. Wang, C. Wei, Y. Xin, X. Xu, J. Yu, Y. Zhang, X. Zhou, Y. Charles, J. Chen, Y. Chen, Y. Du, W. He, Z. Hu, G. Lai, Q. Li, Y. Liu, W. Sun, J. Wang, Y. Wang, Y. Wu, Y. Wu, D. Yang, H. Yang, Y. Yang, Z. Yang, A. Yin, R. Yuan, Y. Zhang, and Z. Zhou, “Kimi-audio technical report.” [Online]. Available: <http://arxiv.org/abs/2504.18425>
- [2] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, “Qwen2-audio technical report.” [Online]. Available: <http://arxiv.org/abs/2407.10759>
- [3] P. S. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks, “AI deception: A survey of examples, risks, and potential solutions,” *Patterns*, vol. 5, no. 5, 2024.
- [4] University of Waterloo, “Watermarks offer no defence against deepfakes,” <https://uwaterloo.ca/news/media/watermarks-offer-no-defense-against-deepfakes>, Jul. 2025, accessed 2025-10-26.
- [5] T. Guardian, “Ceo of world’s biggest ad firm targeted by deepfake scam,” 2024. [Online]. Available: <https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam>
- [6] Forbes, “Fraudsters cloned company director’s voice in \$35 million heist,” 2021. [Online]. Available: <https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/>
- [7] Y. Yang, Y. Kartynnik, Y. Li, J. Tang, X. Li, G. Sung, and M. Grundmann, “Streamvc: Real-time low-latency voice conversion,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.03078>
- [8] D. Milmo, “Company worker in hong kong pays out £20m in deepfake video call scam,” *The Guardian*, 2 2024. [Online]. Available: <https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam>
- [9] INTERPOL, “Beyond illusions: Synthetic media and law enforcement,” INTERPOL, Tech. Rep., 2024. [Online]. Available: [https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS\\_Report\\_2024.pdf](https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_2024.pdf)
- [10] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, “Proactive detection of voice cloning with localized watermarking,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.17264>
- [11] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, “WavMark: Watermarking for audio generation.” [Online]. Available: <http://arxiv.org/abs/2308.12770>
- [12] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, “Detecting voice cloning attacks via timbre watermarking.” [Online]. Available: <http://arxiv.org/abs/2312.03410>
- [13] P. O’Reilly, Z. Jin, J. Su, and B. Pardo, “Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech.” [Online]. Available: <http://arxiv.org/abs/2504.10782>
- [14] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, “Square attack: a query-efficient black-box adversarial attack via random search.” [Online]. Available: <http://arxiv.org/abs/1912.00049>
- [15] H. Liu, M. Guo, Z. Jiang, L. Wang, and N. Z. Gong, “Audiomarkbench: Benchmarking robustness of audio watermarking,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.06979>
- [16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [17] P. Sharma, M. Kumar, H. K. Sharma, and S. M. Biju, “Generative adversarial networks (gans): introduction, taxonomy, variants, limitations, and applications,” *Multimedia tools and applications*, vol. 83, no. 41, pp. 88 811–88 858, 2024.
- [18] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, “Robust audio watermarking using perceptual masking,” *Signal Process.*, vol. 66, no. 3, p. 337–355, May 1998. [Online]. Available: [https://doi.org/10.1016/S0165-1684\(98\)00014-0](https://doi.org/10.1016/S0165-1684(98)00014-0)
- [19] D. Kirovski and H. Malvar, “Spread-spectrum watermarking of audio signals,” *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, 2003.
- [20] M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji, “SilentCipher: Deep audio watermarking,” in *Interspeech 2024*. ISCA, 2024, pp. 2235–2239. [Online]. Available: [https://www.isca-archive.org/interspeech\\_2024/singh24\\_interspeech.html](https://www.isca-archive.org/interspeech_2024/singh24_interspeech.html)
- [21] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. [Online]. Available: <https://arxiv.org/abs/1612.01840>
- [22] D. Gabor, “Theory of communication. part 1: The analysis of information,” *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, pp. 429–441, 1946. [Online]. Available: <https://digital-library.theiet.org/doi/abs/10.1049/ji-3-2.1946.0074>
- [23] M. W. Fakhr, “Robust watermarking using compressed sensing framework with application to mp3 audio,” *The International Journal of Multimedia & Its Applications (IJMA)*, vol. 4, no. 6, pp. 27–43, 2012.
- [24] F. Y. Shih, *Digital watermarking and steganography: fundamentals and techniques*. CRC press, 2017.
- [25] F. H. Hartung, J. K. Su, and B. Girod, “Spread spectrum watermarking: Malicious attacks and counterattacks,” in *Security and Watermarking of Multimedia Contents*, vol. 3657. SPIE, 1999, pp. 147–158.

- [26] J. Dong, X.-J. Mao, C. Shen, and Y.-B. Yang, “Learning deep representations using convolutional auto-encoders with symmetric skip connections,” 2017. [Online]. Available: <https://arxiv.org/abs/1611.09119>
- [27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [28] T. Sainburg, M. Thielk, and T. Q. Gentner, “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires,” *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.
- [29] T. Sainburg, “timsainb/noisereducer: v1.0,” Jun. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3243139>
- [30] H. Foerster, S. Behrouzi, P. Rieger, M. Jadhwal, and A.-R. Sadeghi, “LightShed: Defeating Perturbation-based Image Copyright Protections.”
- [31] “MelScale 2014; TorchAudio 2.8.0 documentation — docs.pytorch.org,” <https://docs.pytorch.org/audio/main/generated/torchaudio.transforms.MelScale.html>, 2025.
- [32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210. [Online]. Available: <http://ieeexplore.ieee.org/document/7178964/>
- [33] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” 2022. [Online]. Available: <https://arxiv.org/abs/2210.13438>
- [34] Z. Wang, J. Guo, J. Zhu, Y. Li, H. Huang, M. Chen, and Z. Tu, “Sleepmarker: Towards robust watermark against fine-tuning text-to-image diffusion models,” *arXiv preprint arXiv:2412.04852*, 2024, focuses on watermarking diffusion models to survive downstream fine-tuning.
- [35] Y. Wen *et al.*, “Tree-ring watermarks: Invisible fingerprints for diffusion model outputs,” in *NeurIPS 2023*, 2023, cited as embedding concentric Fourier-latent patterns in diffusion noise.
- [36] H. Ci, P. Yang, Y. Song, and M. Z. Shou, “Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification,” in *ECCV 2024*, 2024, extends Tree-Ring to multi-key watermark identification.
- [37] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
- [38] R. S. Roman, P. Fernandez, A. Deleforge, Y. Adi, and R. Serizel, “Latent watermarking of audio generative models,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5, ISSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/10889782/>
- [39] H. Guo, B. Chen, Y. Wang, Q. Yan, and L. Xiao, “Audio WaterMark: Dynamic and harmless watermark for black-box voice dataset copyright protection.”
- [40] W. Liu, Y. Li, D. Lin, H. Tian, and H. Li, “GROOT: Generating robust watermark for diffusion-model-based audio synthesis,” in *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, pp. 3294–3302. [Online]. Available: <https://dl.acm.org/doi/10.1145/3664647.3680596>
- [41] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, “A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?” [Online]. Available: <http://arxiv.org/abs/2505.19663>

## Appendix A.

### Comparison Between Watermarked and Watermark-Removed Spectrograms

The results in this section complement the spectrogram analyses discussed in Section 4.3. Instead of plotting the full audio signals, we focus on the watermark residuals before and after removal by HarmonicAttack and AudioSquareAttack. Figure 6 corresponds to Figure 4, and Figure 7 corresponds to Figure 5.

## Appendix B.

### Discussion on Hyperparameter Search

Figure 8 illustrates how ASR varies with different combinations of the loss weights  $\alpha_r$ ,  $\alpha_p$ , and  $\alpha_a$ . For each heatmap, the results are averaged over all values of the third hyperparameter to visualize the pairwise relationships between the remaining two. For this analysis, we fix  $\alpha_{wd}$  to 0.1, because we want to focus on the discussion of psychoacoustic loss and adversarial loss as they are unique to our design. As shown in Figures 8a and 8b, ASR exhibits an approximately linear relationship with  $\alpha_r$  and  $\alpha_p$ : higher ASR is achieved when  $\alpha_r$  is smaller. This trend is intuitive—placing excessive emphasis on the reconstruction loss preserves audio fidelity but weakens watermark removal effectiveness. Figures 8b and 8c further reveal that increasing  $\alpha_a$ , corresponding to the discriminator’s adversarial loss, substantially improves ASR. For instance, when  $\alpha_r = 0.1$ , raising  $\alpha_a$  from 0.01 to 0.5 improves ASR from 0.76 to 1.0 (i.e., perfect removal). Likewise, when  $\alpha_p = 0.001$ , setting  $\alpha_a = 0.5$  achieves perfect ASR. This indicates that adversarial supervision enables the model to more effectively capture and suppress watermark patterns. However, when  $\alpha_a$  becomes dominant, the model tends to overfit to adversarial cues, leading to instability and minor reconstruction artifacts. Finally, Figure 8c shows a nonlinear interaction between the psychoacoustic and adversarial objectives: moderate weighting of both yields the best trade-off between perceptual fidelity and watermark removal. Based on our experiments, the best combination is when  $\alpha_a$  is set to its highest,  $\alpha_p$  is set to its moderate (i.e., balance between quality and removal effectiveness), and  $\alpha_r$  is set to its lowest (i.e.,  $\alpha_a = 0.5$ ,  $\alpha_p = 0.001$  and  $\alpha_r = 0.1$ ). Overall, across all combinations, ASR increases monotonically with  $\alpha_a$ , but decreases with  $\alpha_r$ , demonstrating the inherent trade-off between removal effectiveness and post-removal audio quality.



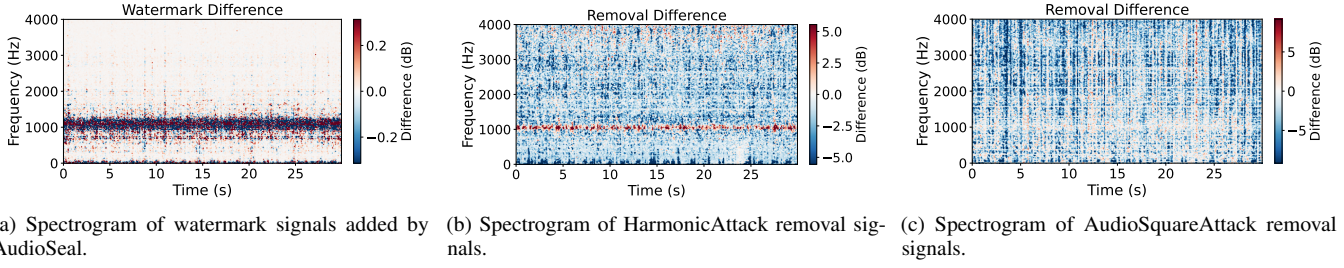


Figure 6. Comparison of watermark signal spectrograms, HarmonicAttack removal spectrograms, and AudioSquareAttack removal spectrograms for AudioSeal-watermarked FMA audio. HarmonicAttack is evaluated by transferring from the model trained on AudioSeal LibriSpeech samples.

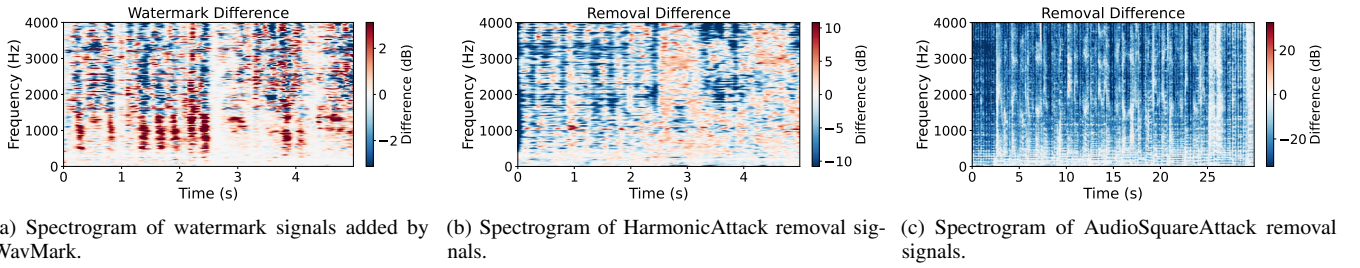


Figure 7. Comparison of watermark signal spectrograms, HarmonicAttack removal spectrograms, and AudioSquareAttack removal spectrograms for WavMark-watermarked FMA audio. HarmonicAttack is evaluated by transferring from the model trained on AudioSeal LibriSpeech samples.

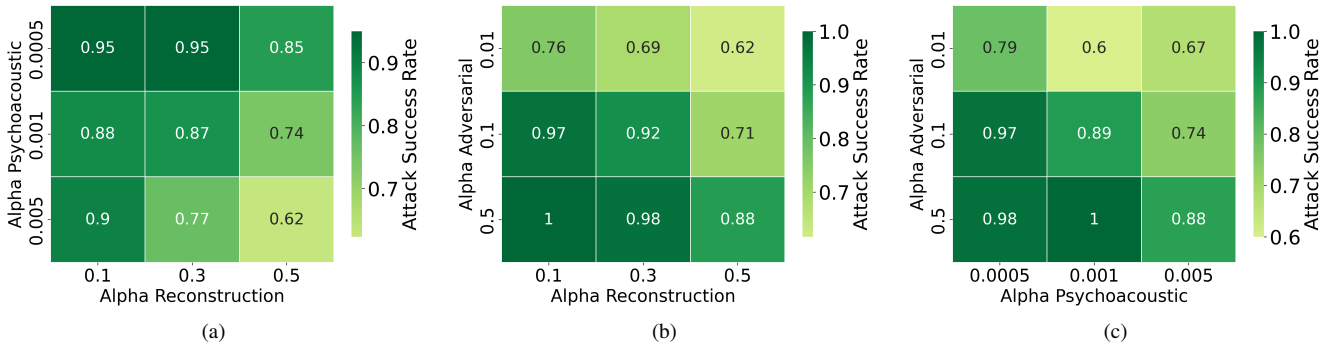


Figure 8. ASR under varying loss-weight combinations across reconstruction ( $\alpha_r$ ), psychoacoustic ( $\alpha_p$ ), and discriminator adversarial ( $\alpha_a$ ) loss components.