Yours or Mine? Overwriting Attacks Against Neural Audio Watermarking

Lingfeng Yao*[1], Chenpei Huang*[1], Shengyao Wang[2], Junpei Xue[2], Hanqing Guo[3], Jiang Liu[2], Phone Lin[4], Tomoaki Ohtsuki[5], Miao Pan[1]

[1] University of Houston
[2] Waseda University
[3] University of Hawaii at Mānoa,
[4] National Taiwan University,
[5] Keio University

lyao12@uh.edu, chuang30@uh.edu, swan@ruri.waseda.jp, akane123@akane.waseda.jp, guohanqi@hawaii.edu, jiang@waseda.jp, plin@csie.ntu.edu.tw, ohtsuki@keio.jp, mpan2@uh.edu

Abstract

As generative audio models are rapidly evolving, AI-generated audios increasingly raise concerns about copyright infringement and misinformation spread. Audio watermarking, as a proactive defense, can embed secret messages into audio for copyright protection and source verification. However, current neural audio watermarking methods focus primarily on the imperceptibility and robustness of watermarking, while ignoring its vulnerability to security attacks. In this paper, we develop a simple yet powerful attack: the overwriting attack that overwrites the legitimate audio watermark with a forged one and makes the original legitimate watermark undetectable. Based on the audio watermarking information that the adversary has, we propose three categories of overwriting attacks, i.e., white-box, gray-box, and black-box attacks. We also thoroughly evaluate the proposed attacks on state-of-the-art neural audio watermarking methods. Experimental results demonstrate that the proposed overwriting attacks can effectively compromise existing watermarking schemes across various settings and achieve a nearly 100% attack success rate. The practicality and effectiveness of the proposed overwriting attacks expose security flaws in existing neural audio watermarking systems, underscoring the need to enhance security in future audio watermarking designs.

Introduction

With the rapid development of generative audio models, artificially producing highly realistic speeches is becoming accessible. This progress also introduces new social risks. For example, attackers can exploit these models to clone or impersonate a person's voice, enabling voice fraud or copyright infringement. These risks arise from a key limitation: human listeners struggle to tell AI-generated audio from real speech audio. To address this problem, audio watermarking has emerged as a proactive defense mechanism. By embedding imperceptible digital signatures into audio signals, watermarking enables future verification of copyright ownership or identification of the audio's source.

To improve usability, existing neural audio watermarking methods primarily focus on two properties: robustness and imperceptibility. Robustness ensures that the watermark remains detectable after common signal processing operations such as compression. Imperceptibility guarantees that watermark embedding does not degrade perceptual audio quality. These objectives have driven recent progress in watermarking research. AudioSeal (Roman et al. 2024) is based on the Encodec (Défossez et al. 2022) architecture, embedding redundant watermark messages in the embedding layer to improve robustness, while introducing novel perceptual loss to preserve audio quality. Timbre (Liu et al. 2024a) embeds the watermark in the frequency domain and incorporates a distortion layer during training to simulate real-world perturbations, thereby maintaining robustness. WavMark (Chen et al. 2024) adopts invertible neural networks (Dinh, Krueger, and Bengio 2015) to embed imperceptible watermarks and use pattern bits to improve robustness against distortions.

Although the robustness of audio watermarking has been extensively studied (Wen et al. 2025; O'Reilly et al. 2025; Özer et al. 2025), its security aspects remain underexplored. Robustness refers to the ability to withstand unintentional perturbations, whereas security concerns its resilience against intentional manipulation by adversaries (Hartung and Kutter 1999; Furon and Duhamel 2003; Li et al. 2021). Roman et al. (2024) was the first to highlight potential security threats in neural audio watermarking. It showed that adversaries, with access to the watermark detector, can launch two adversarial attacks: removal attacks, which make the watermark undetectable, and forgery attacks, which falsely embed a watermark into clean audio. Liu et al. (2024b) extended these attack paradigms and systematically evaluated the vulnerabilities of neural audio watermarking methods under such threats. Furthermore, Liu et al. (2024a) discussed overwriting attacks, demonstrating that an attacker with access to the watermark embedder can insert a new watermark that effectively overwrites the original one. However, their attack relies on a white-box assumption, where the adversary has full knowledge of the watermarking framework, and the evaluation of the overwriting attack lacks thorough analysis.

In this work, we present the first systematic study of wa-

*These authors contributed equally.

termark overwriting attacks, a previously underexplored but practically powerful threat. Unlike removal or forgery attacks, overwriting attacks embed a new watermark to replace the original legitimate one and thus hijack the ownership of the target audio. We perform comprehensive evaluations of three state-of-the-art neural audio watermarking methods (Roman et al. 2024; Liu et al. 2024a; Chen et al. 2024) under the proposed white-box, gray-box, and black-box overwriting attacks. In all threat settings, we achieve a nearly 100% attack success rate in terms of overwriting the original watermark. These findings expose widespread vulnerabilities in existing neural audio watermarking systems and underscore the need to consider security as a primary design objective, alongside robustness and imperceptibility. Our main contributions are summarized as follows.

- We present the first systematic study of overwriting attacks, a powerful yet underexplored security threat in neural audio watermarking. Our work thoroughly analyzes the mechanisms of such attacks and their effects across various threat models.
- Based on the audio watermarking information the adversary has, we propose three-level overwriting attacks, i.e., white-box, gray-box, and black-box attacks, and develop the corresponding attack procedures.
- Through extensive experiments on three state-of-the-art watermarking methods, we find that the proposed overwriting attacks achieve nearly 100% attack success rates. These analyses and experimental results validate that the overwriting attack is a fundamental security challenge to existing neural audio watermarking designs.

Background and Related Work

Principles of Audio Watermarking

Audio watermarking is a technique that embeds information into audio signals without significantly degrading their perceptual quality. A typical audio watermarking system consists of two key components: a **watermark embedder**, which encodes the information into the audio, and a **watermark detector**, which detects and recovers the embedded message. The watermark is typically a fixed-length binary sequence and can carry various types of information depending on the applications. For example, when the watermark encodes copyright metadata, it can support copyright declarations and infringement tracking (Liu et al. 2024a); when it includes source-related tags (e.g., indicators of AI-generated content), it enables source verification (Roman et al. 2024); and when it contains a hash of the audio, it can be used for integrity verification and tampering detection (Yao et al. 2025).

As a proactive protection mechanism, audio watermarking systems typically require the following three properties (Hartung and Kutter 1999): **robustness**, **imperceptibility**, and **security**. Robustness refers to the system's ability to retain the watermark after undergoing common audio processing or transmission (e.g., resampling, compression, or reverberation). Imperceptibility requires that the watermark embedding process introduces no perceptible distortion to audio. Security emphasizes resilience against intentional attacks, ensuring the watermark's integrity even when adversaries attempt to manipulate it.

Neural Audio Watermarking Methods

Audio watermarking has evolved from traditional signal processing techniques into deep learning-based systems, achieving significant progress. Traditional methods, such as least significant bit (LSB) (Cvejic and Seppanen 2004), echo hiding (Gruhl, Lu, and Bender 1996), spread spectrum (Cox et al. 1997), patchwork (Yeo and Kim 2001), and quantization index modulation (QIM) (Chen and Wornell 2001), typically rely on expert knowledge and fixed rules. They are hard to implement and struggle with robustness against complex distortions.

Recent developments in deep learning have introduced a new paradigm for audio watermarking. End-to-end neural watermarking models enable joint optimization of the embedding and detection processes. By leveraging carefully designed loss functions, these models are able to simultaneously enhance robustness and maintain imperceptibility.

Recent neural audio watermarking methods can be broadly categorized into three classes based on their embedding strategies:

- **Encoder-decoder-based approaches**, such as AudioSeal (Roman et al. 2024), XattnMark (Liu et al. 2025), and SilentCipher (Singh et al. 2024), embed watermarks in the high-dimensional latent space learned by neural networks;
- **Frequency-domain-based approaches**, such as Timbre (Liu et al. 2024a) and DeAR (Liu et al. 2023), embed watermarks into the frequency spectrum of audio;
- **Invertible neural network (INN)-based approaches**, such as WavMark (Chen et al. 2024) and IDEAW (Li et al. 2024), model the embedding and detection processes as reversible transformations, enabling high-fidelity and accurate watermark recovery.

In this work, we select three representative systems, AudioSeal, Timbre, and WavMark, from the respective categories. We systematically evaluate their vulnerability to overwriting attacks, a practical and previously underexplored attack in neural audio watermarking.

Security Challenges and Existing Attacks

In audio watermarking systems, security and robustness are two closely related yet fundamentally different properties. Robustness refers to the ability to withstand benign processing operations that are not intended to affect the watermark. In contrast, security concerns its resilience against intentional attacks, where adversaries aim to remove, forge, or manipulate the watermark, and may even possess partial or full access to the watermarking system.

Most existing neural audio watermarking methods prioritize robustness and imperceptibility, while largely overlooking security. Traditional audio watermarking methods (Furon and Duhamel 2003) often leverage the secret key to control the embedding location and detection process,

thus preventing unauthorized manipulation. In contrast, neural watermarking systems typically lack such explicit key-based security mechanisms, but rely on the assumption that the secret of the model weights. However, this assumption is fragile in modern research environments, where open-sourcing and reverse engineering are common practices. According to **Kerckhoffs's principle** (Kerckhoffs 1883), a secure system should remain secure even if everything about the system is public except the secret key. Therefore, relying on "security through obscurity" is inadequate for neural watermarking systems.

Recent studies have begun to explore the security vulnerabilities of neural watermarking. Roman et al. (2024) first demonstrated that adversaries with access to the watermark detector could launch two types of adversarial attacks: removal attacks, which make the watermark undetectable, and forgery attacks, which falsely embed a watermark into clean audio. Liu et al. (2024b) extended this line of work with a more systematic evaluation of neural audio watermarking methods under adversarial attacks.

However, **overwriting attacks**, where adversaries embed a new watermark into an already watermarked audio to override the original ownership, remain largely underexplored. This attack poses a realistic and severe threat to ownership verification. Although Liu et al. (2024a) pointed out overwriting attacks in white-box scenarios, their discussion was limited to insider threats and did not provide a comprehensive evaluation across different adversarial settings. In this work, we address this gap by systematically investigating overwriting attacks across various threat models, detailing their implementations and evaluating their effects on existing neural audio watermarking systems.

System and Threat Model

System Model

Let x denote a clean audio and $m \in \{0,1\}^L$ represent an L-bit binary message intended for embedding. A neural audio watermarking system comprises two key components: an embedder $\mathcal{E}$ and a detector $\mathcal{D}$. The embedder embeds the message into the audio, generating a watermarked signal $x_w = \mathcal{E}(x, m)$. The detector recovers the message $\hat{m}$ from the watermarked audio as $\hat{m} = \mathcal{D}(x_w)$. Ideally, the recovered message matches exactly the embedded message, i.e., $\hat{m} = m$. In practice, a watermarking system must satisfy the following three properties.

Imperceptibility. The watermark embedding process is required to avoid perceptible distortions, which is formally expressed as: $d(x, x_w) \leq \epsilon$, where $d(\cdot,\cdot)$ is a perceptual distance metric and ϵ is an auditory tolerance threshold.

Robustness. The embedded watermark is expected to survive common, unintentional audio transformations $\mathcal{T}(\cdot)$, such as resampling and compression, i.e., $\mathcal{D}(\mathcal{T}(x_w)) = m$.

Security. The watermark is expected to resist adversarial or intentional manipulations $\delta(\cdot)$, which aim to remove, alter, or overwrite the legitimate watermark, satisfying $\mathcal{D}(\delta(x_w)) = m$.

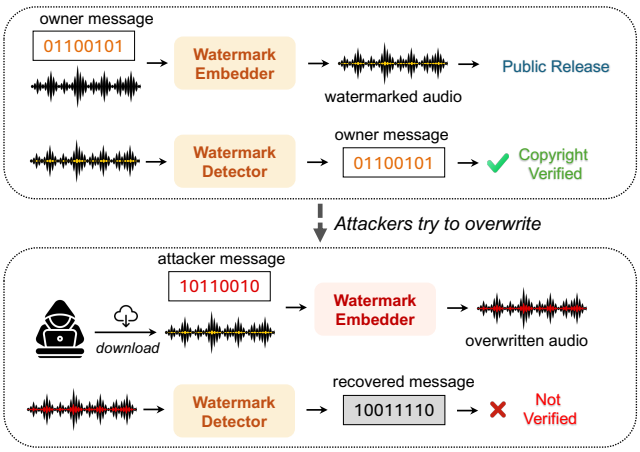


Figure 1: Overview of the proposed audio watermark overwriting attack. An adversary injects a new forged watermark into an already watermarked audio, erasing the original legitimate watermark. Thus, legitimate ownership cannot be verified, and the adversary can falsely claim the copyright.

Previous work focuses on imperceptibility and robustness, while the security property remains underexplored. To bridge this gap, this work systematically investigates an intuitive but practically powerful attack: overwriting attacks on neural audio watermarking systems.

Threat Model

The core threat studied here is the overwriting attack, as shown in Figure 1. The adversary replaces the original watermark with their own, thus hijacking the audio ownership.

Adversarial Goal. To understand the adversarial goal, we first outline the standard copyright verification protocol: a legitimate owner embeds a secret message m_{owner} into audio x, creating a watermarked version $x_w = \mathcal{E}(x, m_{owner})$, which is then publicly distributed. If ownership disputes arise, the owner reveals m_{owner}. Then, the arbiter verifies the ownership by checking whether the detector recovers the matched message, i.e., $\mathcal{D}(x_w) = m_{owner}$.
The adversarial goal is to break this protocol. Given a publicly available watermarked audio x_w, the adversary uses another embedder $\mathcal{E}'$ to embed a new message m'_{adv}, and thus generates a forged audio $x'_w = \mathcal{E}'(x_w, m'_{adv})$. A successful attack satisfies the following conditions: (i) The original message is no longer recoverable: $\mathcal{D}(x'_w) \neq m_{owner}$, (ii) the adversary's detector accurately extracts their own message: $\mathcal{D}'(x'_w) = m'_{adv}$, and (iii) the forged audio remains imperceptible from the original one: $d(x'_w, x_w) \leq \epsilon$.

Attacker Capabilities. Based on the adversary's knowledge of legitimate audio watermarking systems, we propose the following overwriting attacks correspondingly.

- **White-box:** The adversary has full access to the original watermarking embedder $\mathcal{E}$, representing insider threats or attacks on fully open-sourced watermarking models.

- **Gray-box:** The adversary has partial knowledge. They know the general architecture of the watermarking models (e.g., SOTA watermarking designs), but lack the knowledge of model weights and training details. To implement such an attack, the adversary must train a surrogate model $\mathcal{E}'$ to replicate and further replace the functionality of the target watermark.
- **Black-box:** The adversary has no knowledge of the model's architecture or its weights. The black-box overwriting attacks can be classified into two subcategories.
  - **Query-based:** The adversary has limited API access to the original legitimate detector $\mathcal{D}$ output. They can make a few queries to infer the specific watermarking system and then train a surrogate embedder.
  - **Zero-query:** The adversary has no query access to the detector $\mathcal{D}$. They apply a set of public watermarking models or retrained surrogate models to the target watermarked audio speech in a brute-force manner.

Overwriting Attack Designs

In this section, we present the proposed overwriting attack designs in detail. We aim to embed an adversarial message m'_{adv} into an already legitimately watermarked audio x_w and then generate an overwritten audio x'_w. m'_{adv} is an arbitrary binary sequence selected by the attacker to represent forged ownership or other identifying information.

White-box Attack

In the white-box setting, the attacker has full knowledge and access to the original legitimate watermarking embedder. Thus, the overwriting attack is straightforward:

$$x'_w = \mathcal{E}(x_w, m'_{adv}). \tag{1}$$

Gray-box Attack

In the gray-box setting, the adversary lacks access to the weights, training data, and precise training details (e.g., loss functions) of the original watermarking model. Therefore, adversaries must train a surrogate watermarking system $(\mathcal{E}', \mathcal{D}')$. We propose a general watermark training framework to achieve this, which optimizes a joint loss to balance message embedding accuracy and audio imperceptibility.

$$\mathcal{L}_{\text{total}} = \lambda_w \cdot \mathcal{L}_w + \lambda_t \cdot \mathcal{L}_{\text{recon}_t} + \lambda_f \cdot \mathcal{L}_{\text{recon}_f} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}}, \tag{2}$$

where λ_w, λ_t, λ_f, and λ_{adv} are hyperparameters to balance these terms. The individual loss components are as follows.

- **Watermark Recovery Loss ($\mathcal{L}_w$)**: To ensure accurate embedding and detection, we apply binary cross-entropy (BCE) loss between the input message m and the detected message from the surrogate system:

$$\mathcal{L}_{\text{w}} = \text{BCE}(m, \mathcal{D}'(\mathcal{E}'(x, m))). \tag{3}$$

- **Time-domain Reconstruction Loss ($\mathcal{L}_{\text{recon}_t}$)**: To minimize audible distortions, we employ mean squared error (MSE) between the clean and watermarked audio signals:

$$\mathcal{L}_{\text{recon}_t} = \text{MSE}(x, \mathcal{E}'(x, m)). \tag{4}$$

- **Frequency-domain Reconstruction Loss ($\mathcal{L}_{\text{recon}_f}$)**: To further reduce perceptual differences in the frequency domain, we introduce a multi-resolution short-time Fourier transform (STFT) loss (Yamamoto, Song, and Kim 2020). For each resolution m, the loss includes a spectral convergence and a log-magnitude term:

$$\mathcal{L}_{\text{sc}}^{(m)} = \frac{\|S_m(x) - S_m(\mathcal{E}'(x, m))\|_F}{\|S_m(x)\|_F}, \tag{5}$$

$$\mathcal{L}_{\text{mag}}^{(m)} = \frac{1}{N}\|\log(S_m(x)) - \log(S_m(\mathcal{E}'(x, m)))\|_1, \tag{6}$$

where $S_m(\cdot)$ denotes the STFT operation at resolution m, and N is the number of spectrogram elements. The aggregated frequency-domain loss is computed as:

$$\mathcal{L}_{\text{recon}_f} = \frac{1}{M}\sum_{m=1}^{M}\left(\mathcal{L}_{\text{sc}}^{(m)} + \mathcal{L}_{\text{mag}}^{(m)}\right). \tag{7}$$

- **Adversarial Loss ($\mathcal{L}_{\text{adv}}$)**: To further enhance perceptual quality, we employ adversarial training. A discriminator D is trained to distinguish x from $\mathcal{E}'(x, m)$, while the surrogate embedder aims to generate watermarked audio indistinguishable from the original one, i.e.,

$$\mathcal{L}_{\text{d}} = -\log(\sigma(D(x))) - \log(1 - \sigma(D(\mathcal{E}'(x, m)))), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The adversarial loss for the embedder is:

$$\mathcal{L}_{\text{adv}} = -\log(\sigma(D(\mathcal{E}'(x, m)))). \tag{9}$$

After training, the surrogate embedder $\mathcal{E}'$ is ready for overwriting attacks as defined in the white-box scenario.

Black-box Attack

In the black-box setting, adversaries have no knowledge of the original model architecture or its parameters. Attacking strategies vary according to query accessibility:

- **Zero-query Attack:** Without any query access, adversaries collect or reproduce a set of public watermarking models $\mathcal{E}'_i$, and then apply them as much as possible in a brute-force manner to overwrite the original watermark.

$$x_w^{(N)} = (\mathcal{E}_N \circ \mathcal{E}_{N-1} \circ \cdots \circ \mathcal{E}_1)(x_w, m'_{adv}), \tag{10}$$

where $\mathcal{E}_i$ denotes the i-th surrogate watermark embedder collected or trained by adversaries, $x_w^{(N)}$ is the resulting audio after sequentially applying N surrogate embedders, and $\circ$ denotes the function composition operation, which indicates the sequential application from the innermost to the outermost embedder.
- **Query-based Attack:** With limited query access to the original detector $\mathcal{D}$, a more efficient strategy is possible.
  - Partially train the candidate surrogate models for a limited number of epochs.
  - Use these undertrained models to embed new messages into x_w;
  - Query the original detector $\mathcal{D}$ to evaluate whether the original watermark has been tampered;
  - Identify the most effective candidate and refine training until it can reliably perform overwriting attacks.

The query-guided black-box attack significantly reduces computational costs while increasing attacking efficacy.

Experiments and Analyses

To evaluate the vulnerability of neural audio watermarking systems to overwriting attacks, we conduct extensive experiments on three representative methods: AudioSeal (Roman et al. 2024), Timbre (Liu et al. 2024a), and WavMark (Chen et al. 2024). Evaluations are performed under three threat models: white-box, gray-box, and black-box.

Experiment Settings

Datasets and Training Setup. We conduct experiments on two widely used speech datasets: **LibriSpeech** (Panayotov et al. 2015), a corpus of approximately 1,000 hours of English read speech, and **VoxCeleb1** (Nagrani, Chung, and Zisserman 2017), which contains over 150,000 utterances from 1,251 celebrities. All audio samples are resampled to 16kHz and converted to WAV format for consistency across models. All models are trained on a server equipped with 64 CPU cores and two NVIDIA A100 GPUs.

Metrics. We assess attack performance using the following metrics:

- **Bit Error Rate (BER)** measures the proportion of incorrectly recovered bits. Given the embedded message $m \in \{0,1\}^L$ and the detected message $\hat{m}$:

$$\text{BER} = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}\left[m_i \neq \hat{m}_i\right], \tag{11}$$

- **Attack Success Rate (ASR)** reflects how often the embedded message is corrupted after the overwriting attack. It is computed as the proportion of samples where the detected message differs from the original:

$$\text{ASR} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\left[\mathcal{D}(x'_{w,j}) \neq m_j\right], \tag{12}$$

  where $x'_{w,j}$ is the j-th overwritten audio, N is the total number of evaluated samples, and $\mathbb{1}[\cdot]$ is the indicator function.

- **Signal-to-Noise Ratio (SNR)** quantifies the audio distortion introduced by overwriting. It compares the power of the original watermarked audio x_w and the overwritten audio x'_w:

$$\text{SNR} = 10\log_{10}\left(\frac{|x_w|_2^2}{|x_w - x'_w|_2^2}\right), \tag{13}$$

Overwriting Attack Results

We first consider the white-box setting, where the attacker has full access to the original legitimate watermarking model. Figure 2 shows the BER between the original message and the recovered message after being overwritten. The x-axis represents the watermarking method used to embed the original message, and the y-axis denotes the one used by the attacker to embed the overwriting watermark.

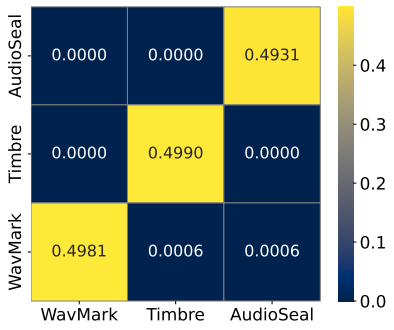


Figure 2: Bit error rate (%) of the original watermark after white-box overwriting.

Metric	Timbre	AudioSeal	WavMark
ASR	99.80	100.00	100.00
ACC	100.00	100.00	100.00

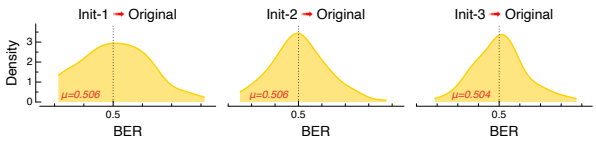
Table 1: White-box overwriting results. Attack success rate (ASR) of the original watermark and recovery accuracy (ACC) of the overwritten watermark (%).

White-box Attack. The diagonal entries (the same method used for both original and overwriting watermarks) consistently reach a BER near 0.5, which means random guessing. This confirms that the original watermark is completely disrupted when the same watermarking method is reused for overwriting. In contrast, off-diagonal values (different methods) show very low BERs, suggesting that overwriting using a different watermarking method fails to destroy the original legitimate watermark. This is because different watermarking methods operate in distinct embedding domains and rely on method-specific decoding mechanisms.

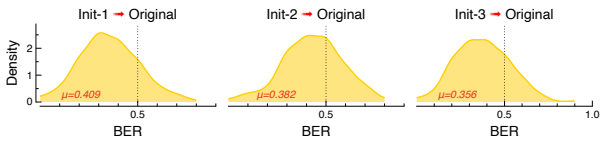
Table 1 summarizes the attack success rate (ASR) of the original watermark and watermark recovery accuracy of the overwritten watermark. All three methods achieve an ASR near 100%, indicating that the original watermark is no longer verifiable. Meanwhile, the overwritten watermark is recovered with perfect accuracy, indicating that the attacker has successfully hijacked the audio ownership.

These results reveal a critical vulnerability: when the attackers have access to the same watermarking embedder, they can reliably overwrite the original legitimate watermark with their own, effectively hijacking the audio ownership.

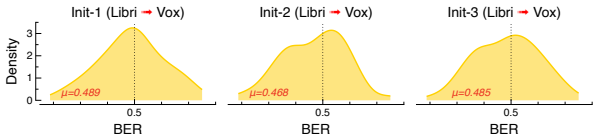
Gray-box Attack. In the gray-box setting, the adversary knows the watermarking system's architecture but lacks access to its training details or training data. To assess this scenario, we construct surrogate models under two settings: (i) cross-training, where the surrogate model is trained on the same dataset (VoxCeleb1) but with a different training pipeline, and (ii) cross-data, where the surrogate is trained on an entirely different dataset (LibriSpeech). We train three surrogate models with varying random seeds, denoted as Init-1, Init-2, and Init-3, respectively. The "Init-k→Official" notation indicates that surrogate Init-k attempts to overwrite the watermark embedded by official model.



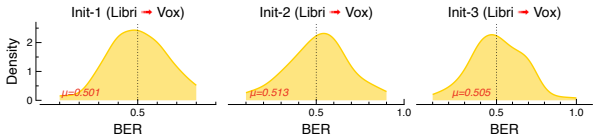
(a) BER of AudioSeal (cross-training).



(b) BER of Timbre (cross-training).



(c) BER of on AudioSeal (cross-data).



(d) BER of Timbre (cross-data).

Figure 3: Bit error rate (%) distributions of the original watermark for AudioSeal and Timbre under gray-box settings.

Method	Init-1	Init-2	Init-3
ASR (Original Watermark)			
Timbre	99.60	98.80	98.40
AudioSeal	100.00	100.00	100.00
WavMark	100.00	100.00	99.50
ACC (Overwritten Watermark)			
Timbre	100.00	100.00	100.00
AudioSeal	99.40	99.00	99.80
WavMark	100.00	100.00	100.00

Table 2: Gray-box cross-training results (%). Top: attack success rate (ASR) of the original watermark. Bottom: recovery accuracy (ACC) of the overwritten watermark.

Method (Libri → Vox)	Init-1	Init-2	Init-3
ASR (Original Watermark)			
Timbre	99.80	99.90	98.80
AudioSeal	100.00	100.00	100.00
WavMark	100.00	100.00	100.00
ACC (Overwritten Watermark)			
Timbre	99.90	100.00	99.90
AudioSeal	97.20	98.90	99.50
WavMark	100.00	100.00	100.00

Table 3: Gray-box cross-data results (%). Top: attack success rate (ASR) of the original watermark. Bottom: recovery accuracy (ACC) of the overwritten watermark.

Figures 3a and 3b present the bit error rate (BER) distributions of the original watermark under cross-training attacks. For AudioSeal, all surrogate models yield BER distributions tightly centered around 0.5 (mean μ = 0.504 - 0.506). This indicates a complete corruption of the original watermark, as the detection performance is consistent with random guessing. In contrast, attacks on Timbre yielded distributions around 0.4, suggesting that while most bits of the watermark are corrupted, some residual information still remains. We exclude WavMark from the BER distribution analyses because its decoder uses pattern bit verification: once the watermark is overwritten, the pattern check fails and the decoder outputs nothing, which makes BER impossible to define. Therefore, we report only ASR and recovery accuracy for WavMark.

The cross-training quantitative results are summarized in Table 2. For all watermarking methods, the attack success rate approaches 100%, demonstrating that surrogate models consistently invalidate the original watermark. Simultaneously, the overwritten watermark is recovered with near-perfect accuracy. This suggests that different training configurations do not significantly alter the embedding behavior to resist overwriting attacks.

In cross-data scenarios, Figures 3c and 3d illustrate that surrogate models trained on LibriSpeech effectively disrupt watermarks embedded by models trained on VoxCeleb1.

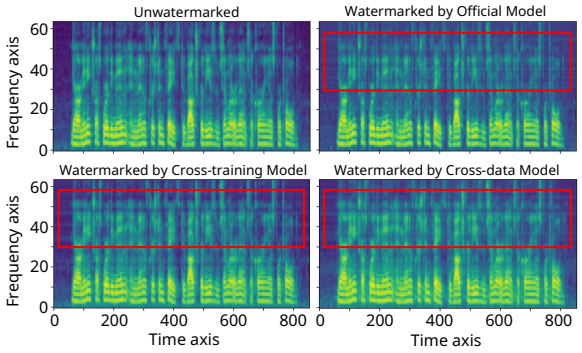
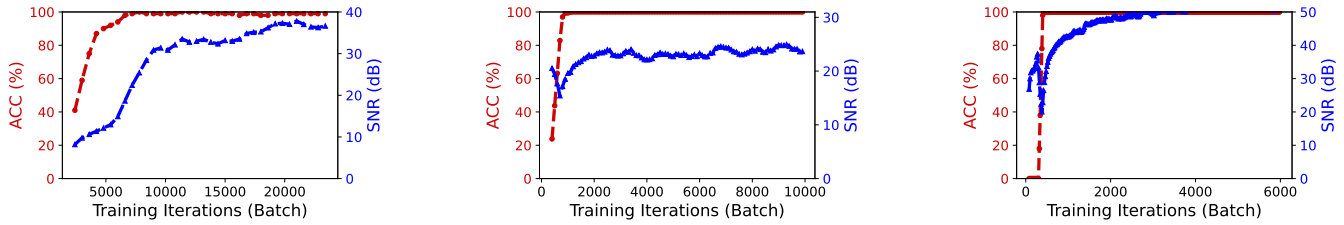


Figure 4: Spectrogram comparison of the unwatermarked audio and three watermarked versions. Red boxes highlight the spectral perturbations introduced by the watermark.

The resulting BER distributions are centered around 0.5 for both AudioSeal and Timbre, indicating complete erasure of the original legitimate watermark.

Table 3 quantitatively supports these observations, showing a near-perfect ASR and watermark recovery accuracy in all methods. The results demonstrate that surrogate models trained on varying datasets can effectively overwrite and replace original legitimate watermarks.

We further illustrate the spectrograms of unwatermarked



(a) AudioSeal Reproduction (b) Timbre Reproduction (c) WavMark Reproduction

Figure 5: Black-box reproduction attacks on three watermarking systems.

and watermarked audio versions produced by official, cross-training, and cross-data Timbre models in Figure 4. Despite differences in training details or datasets, all models embed watermarks within similar spectral regions (highlighted in red boxes). This consistent embedding behavior facilitates successful watermark overwriting, challenging the security assumption that secret training details or private weights are sufficient to ensure the watermark security.

Black-box Attack. In the black-box scenario, the adversary has no prior knowledge of the watermarking algorithm, model weights, or training data. We analyze two practical strategies: zero-query and query-based attacks.

Under the zero-query attack, the adversary collects or reproduces a set of candidate watermarking models and applies them sequentially in a brute-force manner to overwrite the original watermark. Figure 6 illustrates the impact of progressively stacking multiple watermarking methods, assuming a scenario with three common watermarking techniques. As more watermarking models are sequentially applied, the overwriting ASR increases from nearly 30% (one embedder) to almost 100% (three embedders), while SNR decreases from about 24 dB to 20 dB. Practically, as the candidate methods expand, the zero-query strategy becomes increasingly inefficient, leading to significant perceptual degradation and high computational overhead.

The query-based attack mitigates these drawbacks by utilizing limited queries to the watermark detector. Instead of exhaustively applying all candidates, the adversary iteratively "try-and-test" candidates and stops once the detector confirms successful overwriting of the original legitimate watermark. Consequently, audio degradation is limited to a single embedding operation. As depicted in Figure 5, watermarking models exhibit varying convergence time. Certain methods reach sufficient overwriting capabilities early in the training, allowing effective overwriting with partially trained embedders. Table 4 provides a comparative analysis between zero-query and query-based attacks. The results indicate that query-based attacks, using fewer than 10 detector queries, achieve a reduction of over 50% in training iterations compared to the zero-query approach. By applying only a single effective watermarking method instead of stacking multiple methods, query-based attacks preserve the audio quality and maintain identical attack success rates. As the size of the candidate set grows, the advantages of the query-based approach become more pronounced.

Attack Type	Query	Training Cost	SNR (dB)	ASR (%)
Zero-query	0	36,000 iters	20.63	100
Query-based	<10	14,000 iters	24.19	100

Table 4: Comparison of zero-query and query-based black-box overwriting attacks.

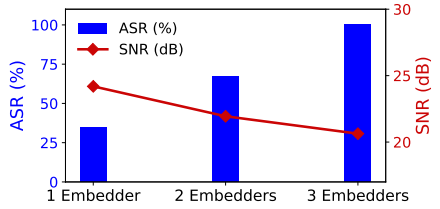


Figure 6: Black-box attack success rate and audio quality.

Discussion and Conclusion

This paper presents the first systematic study of overwriting attacks, uncovering a critical security vulnerability in existing neural audio watermarking paradigms. Our results demonstrate that watermarking methods, while effectively optimized for robustness and imperceptibility, are vulnerable to intentional overwriting attacks. Consequently, audio ownership can be readily hijacked, significantly undermining watermarking's reliability for provenance verification.

The effectiveness of white-box attacks highlights a critical oversight: existing neural audio watermarking methods rely on model secrecy for security, neglecting explicit defenses against intentional manipulation. Our gray-box experiments further reveal that surrogate models, even trained on different data and implementation details, can consistently converge to similar embedding strategies. This architectural convergence undermines the assumption that the secrecy of model weights alone suffices for security. Additionally, our black-box evaluation confirms the practicality and feasibility of overwriting attacks under minimal knowledge and limited query access.

In conclusion, these insights highlight the necessity for a fundamental shift in neural watermarking research. Future neural audio watermarking methods must incorporate explicit mechanisms for security and integrate robust defenses against intentional adversarial attacks alongside imperceptibility and robustness goals.

Acknowledgements

The work of L. Yao, C. Huang, and M. Pan was supported in part by the US National Science Foundation under grants CNS-2107057, CNS-2318664, CSR-2403249, and CNS-2431596. The work of P. Lin was supported in part by the National Science and Technology Council (NSTC) of Taiwan under grants NSTC 114- 2221-E-002-141-MY3, NSTC 112-2221-E-002-163-MY3, and NSTC 113-2314-B-A49-034-MY3. The work of T. Ohtsuki was supported in part by JST ASPIRE Grant Number JPMJAP2326, Japan.

References

Chen, B.; and Wornell, G. 2001. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4): 1423–1443.

Chen, G.; Wu, Y.; Liu, S.; Liu, T.; Du, X.; and Wei, F. 2024. WavMark: Watermarking for Audio Generation. arXiv:2308.12770.

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

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

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Non-linear Independent Components Estimation. arXiv:1410.8516.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2022. High Fidelity Neural Audio Compression. arXiv:2210.13438.

Furon, T.; and Duhamel, P. 2003. An asymmetric watermarking method. *IEEE Transactions on Signal Processing*, 51(4): 981–995.

Gruhl, D.; Lu, A.; and Bender, W. 1996. Echo hiding. In *Information Hiding*, 295–315. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-49589-5.

Hartung, F.; and Kutter, M. 1999. Multimedia watermarking techniques. *Proceedings of the IEEE*, 87(7): 1079–1107.

Kerckhoffs, A. 1883. La cryptographie militaire. *J. Sci. Militaires*, 9(4): 5–38.

Li, L.; Shi, D.; Hou, R.; Li, H.; Pan, M.; and Han, Z. 2021. To Talk or to Work: Flexible Communication Compression for Energy Efficient Federated Learning over Heterogeneous Mobile Edge Devices. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 1–10.

Li, P.; Zhang, X.; Xiao, J.; and Wang, J. 2024. IDEAW: Robust Neural Audio Watermarking with Invertible Dual-Embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4500–4511.

Liu, C.; Zhang, J.; Fang, H.; Ma, Z.; Zhang, W.; and Yu, N. 2023. Dear: A deep-learning-based audio re-recording resilient watermarking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 13201–13209.

Liu, C.; Zhang, J.; Zhang, T.; Yang, X.; Zhang, W.; and Yu, N. 2024a. Detecting Voice Cloning Attacks via Timbre Watermarking. In *Network and Distributed System Security Symposium*.

Liu, H.; Guo, M.; Jiang, Z.; Wang, L.; and Gong, N. Z. 2024b. AudioMarkBench: Benchmarking Robustness of Audio Watermarking. In *Advances in Neural Information Processing Systems*, volume 37, 52241–52265.

Liu, Y.; Lu, L.; Jin, J.; Sun, L.; and Fanelli, A. 2025. XAttnMark: Learning Robust Audio Watermarking with Cross-Attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 38987–39015.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, 2616–2620.

O'Reilly, P.; Jin, Z.; Su, J.; and Pardo, B. 2025. Deep Audio Watermarks are Shallow: Limitations of Post-Hoc Watermarking Techniques for Speech. In *The 1st Workshop on GenAI Watermarking*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Roman, R. S.; Fernandez, P.; Elsahar, H.; Défossez, A.; Furon, T.; and Tran, T. 2024. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*.

Singh, M. K.; Takahashi, N.; Liao, W.; and Mitsufuji, Y. 2024. SilentCipher: Deep Audio Watermarking. In *Proc. Interspeech 2024*, 2235–2239.

Wen, Y.; Innuganti, A.; Ramos, A. B.; Guo, H.; and Yan, Q. 2025. SoK: How Robust is Audio Watermarking in Generative AI models? arXiv:2503.19176.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6199–6203.

Yao, L.; Huang, C.; Wang, S.; Xue, J.; Guo, H.; Liu, J.; Chen, X.; and Pan, M. 2025. SpeechVerifier: Robust Acoustic Fingerprint against Tampering Attacks via Watermarking. arXiv:2505.23821.

Yeo, I.-K.; and Kim, H. J. 2001. Modified Patchwork Algorithm: a novel audio watermarking scheme. In *Proceedings International Conference on Information Technology: Coding and Computing*, 237–242.

Özer, Y.; Choi, W.; Serrà, J.; Singh, M. K.; Liao, W.-H.; and Mitsufuji, Y. 2025. A Comprehensive Real-World Assessment of Audio Watermarking Algorithms: Will They Survive Neural Codecs? arXiv:2505.19663.