RANDOMIZED REPRESENTATION SMOOTHING FOR AUDIO WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

With the rapid development of audio generators, audio watermarking has become an important technique for content authentication and provenance verification. In this work, we propose RRS-AW, a method for embedding watermarks into randomized transformation-averaged robust representations of audio during generation. The method encodes binary messages through pairwise coordinate ordering in representations averaged over random audio transformations and optimizes the generation latent while keeping both the generator and watermark extractor fixed. We provide a theoretical analysis of the robustness of the resulting representation and complement it with an empirical analysis of its local sensitivity. Experiments on AudioLDM-generated music demonstrate strong robustness across individual, composite, and codec transformations: RRS-AW achieves the best or tied-best bit accuracy on 9 of 13 composite attacks and 9 of 13 codec settings among the competitors. Additional experiments demonstrate robustness across different audio generators, low false-positive detection, and high perceptual audio quality.

1 INTRODUCTION

Recent advances in generative models enable the synthesis of diverse, high-fidelity music from text prompts. Systems such as MusicGen (Copet et al., 2023) and Suno (Freyberg, 2024) make music generation increasingly accessible, but also introduce risks such as royalty manipulation and audio deepfakes. As human listeners increasingly struggle to distinguish generated from real audio (Yao et al., 2026), reliable identification of generated content becomes important. Audio watermarking addresses this problem by associating generated audio with an embedded message and key for subsequent verification.

Audio watermarking embeds an imperceptible signal into generated audio to facilitate provenance tracking and attribution, while balancing two main objectives: robustness to transformations encountered during transmission or processing, and the imperceptibility of the embedded signal. Existing approaches employ different embedding strategies: AudioSeal (San Roman et al., 2024) introduces a perceptual-quality loss, WavMark (Chen et al., 2023) uses an invertible network, SilentCipher (Singh et al., 2024) operates in a learned latent space, and XattnMark (Liu et al., 2025) uses shared generator-detector components and cross-attention for message retrieval.

Despite significant progress (Salah et al., 2025; Rong & Hu, 2026), the robustness of existing methods is typically established only empirically (Kovačević et al., 2026; Sedaghati et al., 2026). In this work, we take a step towards more robust and theoretically grounded audio watermarking and propose *RRS-AW*, a method that embeds watermarks into randomized transformation-averaged robust representations of audio during its generation. An auxiliary encoder computes a representation averaged over a specified distribution of synthetic perturbations, and a binary message is embedded into secret components of this representation while preserving audio quality. This design improves watermark stability under transformations potentially aimed at watermark removal. Extensive experiments demonstrate robustness to individual, composite, and codec transformations.

Our contributions are summarized as follows:

- We propose *RRS-AW*, an in-generation audio watermarking framework that encodes binary messages through pairwise coordinate ordering in randomized transformation-averaged au-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dio representations. The averaging explicitly incorporates a user-defined distribution of audio transformations into the watermark representation.

- We develop a theoretically motivated robustness formulation for the proposed representation and analyze its sensitivity to input transformations. RRS-AW keeps both the audio generator and watermark extractor fixed and optimizes only the generation latent, avoiding retraining of either model and enabling application to generators with a differentiable latent-to-waveform path.
- We extensively evaluate RRS-AW under individual, composite, and codec transformations, as well as across multiple audio generators. RRS-AW achieves the best or tied-best bit accuracy on 9 of 13 composite attacks and 9 of 13 codec settings among the main baselines, while maintaining high perceptual audio quality and a zero empirical false-positive rate on the evaluated negative set. Ablation experiments further demonstrate the contribution of randomized representation averaging to robustness. The source code is anonymously available at `https://anonymous.4open.science/r/RRS4AW`.

2 Related Work

Digital audio watermarking aims to embed information into an audio signal while preserving perceptual quality and ensuring robustness to signal transformations. Classical approaches include spread-spectrum watermarking, quantization index modulation, patchwork methods, and masking-based techniques (Kirovski & Malvar, 2003; Chen & Wornell, 2001; Yeo & Kim, 2003; Swanson et al., 1998). Although these methods are computationally efficient and interpretable, they usually rely on hand-crafted signal assumptions and manually designed embedding rules. Their robustness, therefore, depends on the expected distortion model and may degrade under different transformations, including strong filtering, resampling, compression, or neural-codec processing.

Recent neural watermarking methods address these limitations by training watermark embedders and detectors with simulated distortions. WavMark (Chen et al., 2023) embeds multi-bit messages into short audio segments using a neural framework designed for generated audio. AudioSeal (San Roman et al., 2024) proposes localized watermark detection and uses a perceptual loss to preserve audio quality. SilentCipher (Singh et al., 2024) introduces a high-fidelity deep watermarking system with psychoacoustic constraints and differentiable compression approximations. Other neural methods target more specific robustness or security settings, such as timbre watermarking for voice-cloning detection, re-recording resilience, adversarial resistance to edits, key-controlled decoding, and codec-integrated watermarking (Liu et al., 2024a; 2023a; Pavlović et al., 2025; Xu et al., 2025; Zhou et al., 2024a). These works establish important baselines for robust audio watermarking.

Another line of work studies watermarking directly in generative, latent, or codec-based audio pipelines. GROOT (Liu et al., 2024c) embeds watermarks during diffusion-based audio generation, while TraceableSpeech (Zhou et al., 2024b) directly generates watermarked speech in a text-to-speech system. VocBulwark (Liu et al., 2026) introduces additional trainable parameters for generative speech watermarking, and XAttnMark (Liu et al., 2025) uses generator-detector parameter sharing and cross-attention for message retrieval. Related latent-space approaches, such as Latent-Mark (Chen et al., 2026), aim to improve robustness to neural resynthesis by targeting representations that are more stable under codec transformations. Such approaches benefit from access to internal representations of the synthesis or codec pipeline. RRS-AW keeps the pretrained generator fixed and embeds the watermark through latent optimization of each generated sample. Our approach combines transformation-averaged representations with pairwise message encoding and sample-specific latent optimization.

3 Methodology

The proposed method combines in-generation watermark insertion, attack-robust representation aggregation, and secret-key-based multi-bit extraction. Its main distinction from post-hoc neural watermarking methods is that watermark recovery is performed from an aggregated representation over multiple transformed samples of the target audio. In addition, unlike generator-specific methods that require retraining or modifying the generator, the proposed approach keeps the pretrained generator intact and embeds the watermark through sample-specific latent optimization.

3.1 PROBLEM STATEMENT AND ASSUMPTIONS

Let $\mathcal{G}$ be a well-trained audio generative model that maps an input prompt to the audio waveform s. The audio waveform is represented by a sequence of floating-point numbers $s_i \in [-1, 1]$. In this work, we consider detection of an embedded message and its association with a registered watermark and key. To do so, we apply in-generation watermarking of the content produced by $\mathcal{G}$. Mathematically, let $e_{\mathcal{G}} : \{0,1\}^n \to [-1,1]^m$ be the function that embeds a binary watermark of length n into the waveform of length m, and $d : [-1,1]^m \to \{0,1\}^n$ be the function that retrieves a binary message of length n from the waveform of length m. Here, we indicate that the watermarking mechanism e depends on the generator $\mathcal{G}$. Below, we omit the subscript to ease the notation. The notation suppresses the prompt, generator randomness, secret key, and transformation samples. The embedding and extraction procedures are designed to satisfy the following objectives:

1. The watermarks are correctly extracted from the generated audio, namely,
$$w = d(e(w)), \tag{1}$$
where $w \in W$ is an embedded watermark, and W is a nonempty set of binary messages of length n used for watermarking.
2. The binary messages extracted from non-generated audio should be far, in Hamming distance, from every registered watermark, namely that for all $w \in W$, where $\Delta = \|w - d(y)\|_1$ and y is the audio waveform *not* generated by $\mathcal{G}$. the following holds: $0 \ll \Delta \ll n$. In other words, the detector should rarely accept unwatermarked audio. Negative examples include both natural recordings and unwatermarked outputs of $\mathcal{G}$ or other generators. This objective is evaluated by the false-positive rate for a specified negative distribution and decision rule.
3. The watermarking process does not introduce a significant quality degradation of the audio. Namely, we want the following inequality to be satisfied for the watermarked audio $e(w)$:
$$\rho(s, e(w)) < \varepsilon. \tag{2}$$
Here, $\rho : [-1,1]^m \times [-1,1]^m \to \mathbb{R}_+$ is the fixed audio quality metric and $\varepsilon > 0$ is a constant.

With these objectives, we want to design such watermark embedding (e) and extraction (d) models that are sufficiently robust to corruptions that s may undergo. Such perturbations include, but are not limited to, parametric transformations (such as bandpass filter or additive noise) or codec-based waveform compression. We model such corruptions by applying audio transformation function $\phi : [-1,1]^m \times \Theta \to [-1,1]^m$, where Θ is the predefined set of parameters. Informally, the goal of the application of ϕ to the watermarked audio s is to force the watermark detector not to extract a watermark from the perturbed watermarked audio.

3.2 DESCRIPTION OF RRS-AW

Embedding of Watermarks Let $(w, \mathbb{I})$ be the pair of the watermark and secret key assigned to the user of the audio generative model, respectively. Here, $w \in \{0,1\}^n$ is the n-bit binary watermark, and $\mathbb{I} = \{(l_i, u_i)\}_{i=1}^n$ is a set of n distinct index pairs. Let $s \in [-1,1]^m$ denote an audio. We embed the watermark into a robust representation of s with respect to the transformation function ϕ. As such a robust representation, we use the expected value
$$\gamma(s) = \mathbb{E}_{\theta \sim \mathcal{U}(\Theta)} g[\phi(s, \theta)], \tag{3}$$
where $\theta \sim \mathcal{U}(\Theta)$ denotes the uniform sampling of parameter θ of transformation ϕ from predefined parameter space Θ, and $g : [-1,1]^m \to \mathbb{R}^d$ is an auxiliary pretrained encoder network. We indicate that the true value of $\gamma(s)$ is intractable, since the distribution of the random variable $g[\phi(s, \theta)]$ is unknown. Thus, instead of the true value of $\gamma(s)$, we use its unbiased estimate in the form
$$\hat{\gamma}(s) = \frac{1}{k} \sum_{j=1}^{k} g[\phi(s, \theta_j)], \tag{4}$$

Algorithm 1 RRS-AW watermark embedding

Require: Prompt p, generator $\mathcal{G}$, its acoustic and vocoder subpart $\mathcal{V}$, watermark w, secret key $\mathbb{I}$, encoder g, transformation families ϕ, sample size k, maximum steps R
Ensure: Watermarked audio $\hat{s}$

Generate non-watermarked audio $s \leftarrow \mathcal{G}(p)$
Initialize initial latent state $z \leftarrow z_0$
for $r = 1, \ldots, R$ **do**
  Decode current latent: $\hat{s} \leftarrow \mathcal{V}(z)$
  Sample $\theta_1, \ldots, \theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\Theta)$ for $\hat{s}$ augmentation
  Estimate the robust representation $\hat{\gamma}(\hat{s})$ as in Eq. equation 4
  Normalize the nonzero aggregate: $h(\hat{s}) \leftarrow \hat{\gamma}(\hat{s})/\|\hat{\gamma}(\hat{s})\|_2$
  Compute $\mathcal{L}_{\text{wm}}(h(\hat{s}), w, \mathbb{I})$, $\mathcal{L}_{\text{qual}}(s, \hat{s})$, and $\mathcal{L}_{\text{total}}$
  if StoppingCriteria$(\mathcal{L}_{\text{wm}}, \mathcal{L}_{\text{qual}})$ is satisfied **then**
    break
  end if
  Update z by one gradient step minimizing $\mathcal{L}_{\text{total}}$
end for
return $\hat{s} = \mathcal{V}(z)$

Algorithm 2 RRS-AW watermark extraction

Require: Query audio y, secret key $\mathbb{I} = \{(l_i, u_i)\}_{i=1}^n$, encoder g, transformation family ϕ, sample size k
Ensure: Extracted watermark $\hat{w}$

Sample $\theta_1, \ldots, \theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\Theta)$
Estimate the robust representation $\hat{\gamma}(y)$ as in Eq. equation 4
Normalize the nonzero aggregate: $h(y) \leftarrow \hat{\gamma}(y)/\|\hat{\gamma}(y)\|_2$
for $i = 1, \ldots, n$ **do**
  Extract $\hat{w}_i$ by comparing $\hat{\gamma}(y)_{l_i}$ and $\hat{\gamma}(y)_{u_i}$ as in Eq. equation 7
end for
return $\hat{w}$

where θ_j are i.i.d. samples from $\mathcal{U}(\Theta)$ and k is the used sample size of transformed audio. Let $[x]_\alpha^+ = \max\{\alpha, x\}$. We embed the watermark by sparsely changing the vector $\hat{\gamma}(s)$ in the positions encoded by the secret key $\mathbb{I}$ by minimizing the following loss function:

$$\mathcal{L}_{\text{wm}}(\hat{\gamma}(s), w, \mathbb{I}) = \frac{1}{n} \sum_{i=1}^{n} \left[\delta - (-1)^{w_i} \left(\hat{\gamma}(s)_{l_i} - \hat{\gamma}(s)_{u_i}\right)\right]_0^+, \tag{5}$$

where $\delta > 0$ is the scalar margin parameter.

In addition to the watermark embedding loss $\mathcal{L}_{\text{wm}}$, a regularization loss $\mathcal{L}_{\text{qual}}$ is added to preserve the quality of generated audio and the imperceptibility of the watermark. Possible choices are, for example, the mean ℓ_2 or ℓ_1 distance over audio amplitudes or spectrogram coefficients. The total loss function combines these two objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{wm}} \mathcal{L}_{\text{wm}} + \lambda_{\text{qual}} \mathcal{L}_{\text{qual}}, \tag{6}$$

where $\lambda_{\text{wm}} > 0$ and $\lambda_{\text{qual}} > 0$ are the fixed constants. The overview of the watermark embedding process is presented in Algorithm 1.

Extraction of Watermark The proposed encoding strategy is paired with a simple decoding rule. Namely, given the query audio s, we compute the representation $\hat{\gamma}(s)$ according to Eq. equation 4 and, given the secret key $\mathbb{I} = \{(l_i, u_i)\}_{i=1}^n$ extract the binary message $\hat{w}$ according to the following scheme:

$$\begin{cases} \hat{\gamma}(s)_{l_i} \geq \hat{\gamma}(s)_{u_i} \implies \hat{w}_i = 0, \\ \hat{\gamma}(s)_{l_i} < \hat{\gamma}(s)_{u_i} \implies \hat{w}_i = 1. \end{cases} \tag{7}$$

Informally, the watermark is encoded in the relations in the secret pairs of indices of the random vector $\hat{\gamma}(s)$. See Algorithm 2 for details.

Robustness Properties In this section, we theoretically analyze the robustness of the proposed scheme to parametric audio perturbations. In our setting, the robustness of the method to audio

perturbations scales with the quantity

$$\Delta(t) = \hat{\gamma}_{\text{enc}}(s) - \hat{\gamma}_{\text{dec}}(\phi(s,t)), \tag{8}$$

where $\phi(s,t)$ is the perturbation that (watermarked) audio s may undergo. For simplicity of derivations, we assume that the perturbation type here is the same as the one used during the watermark embedding (see Eq. equation 3), and t is an unknown parameter. Here, we distinguish $\hat{\gamma}_{\text{enc}}$ from $\hat{\gamma}_{\text{dec}}$ to indicate that they are computed over different samples of noise (see Eq. 4).

The lemma below provides an upper bound for the value from Eq. 8 that holds with high probability.

Lemma 1. *Let $\theta_j \sim \mathcal{U}(\Theta)$ be i.i.d. random variables that represent parameters of parametric perturbation ϕ. Let $X_j(t) = g(\phi(s, \theta_j^{enc})) - g(\phi(\phi(s,t), \theta_j^{dec})) \in \mathbb{R}^d$ and let $m(t) = \mathbb{E}(X_j(t))$. Let $Y_j(t) = X_j(t) - m(t)$. Assume that $\|Y_j(t)\|_2 \leq \mu$ and $\mathbb{E}\|Y_j(t)\|_2^2 \leq \sigma^2$.*

Let $\theta \mapsto g[\phi(s,\theta)]$ be a measurable function. Let the function $g \circ \phi(\cdot, \theta)$ be $L-$Lipschitz and let $\|\phi(s,t) - \phi(s,t_0)\| \leq G\|t - t_0\|_q$, where $\phi(s,t_0) = s$ for all s. Then, with the probability at least $1-\eta$,

$$\|\Delta(t)\|_2 \leq LG\|t - t_0\|_q + \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}} \tag{9}$$

and $\|\cdot\|_q$ is the norm on Θ.

Lemma 2. *Let the watermark w be embedded in the audio $\hat{\gamma}(s)$ with the use of the secret $\mathbb{I} = \{(l_i, u_i)\}_{i=1}^n$. Assume that the watermark embedding optimization converges such that $\mathcal{L}_{wm} = 0$. Let $\Delta_i = |\hat{\gamma}(s)_{l_i} - \hat{\gamma}(s)_{u_i}|$ be the embedding margin for the $i-$th bit. Then, the number of bit flips N_{err} under a perturbation with the parameter t is bounded:*

$$N_{err} \leq |\{i \in \{1, \ldots, n\} : \Delta_i \leq 2\|\Delta(t)\|_\infty\}|. \tag{10}$$

Specifically, with probability at least $1-\eta$ no bit flips occur under perturbation with parameter t such that

$$LG\|t - t_0\|_q + \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}} < \frac{\delta}{2}, \tag{11}$$

where δ is the margin parameter from Eq. 5.

Remark 1. *The lemmas above establish a connection between a perturbation $t - t_0$ in the parameter space Θ and the number of bit flips in the watermark under the corresponding perturbation. The proofs of the lemmas are deferred to the appendix.*

4 Experimental settings

4.1 Base model and evaluation data

We use AudioLDM1 (Liu et al., 2023b) as the base model, a text-to-audio latent diffusion model capable of music synthesis. A pretrained contrastive language-audio pretraining (CLAP) (Elizalde et al., 2023) model maps text and audio into a shared embedding space: during latent diffusion model (LDM) training, the model is conditioned on CLAP audio embeddings, while at inference, these are replaced by CLAP text embeddings. The diffusion process itself does not operate on raw audio, but on a compact variational autoencoder (VAE) latent representation of the mel-spectrogram. Starting from Gaussian noise, the reverse diffusion network iteratively denoises this latent state in the VAE latent space. This generated latent representation is then passed to the VAE decoder, which reconstructs a mel-spectrogram. Finally, a HiFi-GAN (Kong et al., 2020) vocoder converts the reconstructed mel-spectrogram into the time-domain waveform.

For the watermarking experiments in the music generation setting, we used a set of 50 predefined prompts. One example prompt is: “Clear techno music with a strong, upbeat tempo“. Using these prompts, audio with and without watermarks was generated to evaluate the quality of watermark detection, false positive detection of watermarks, and the imperceptibility of the watermark embedding. The negative-set experiments use 512 recordings from the Freesound subset of LAION-Audio-630K (Wu et al., 2023).

4.2 PRIMARY METRIC

The primary metric for the evaluation is bit accuracy (BitAcc, BAC), which can be expressed in terms of ℓ_1 norm or Hamming distance as

$$\mathrm{BitAcc}(w, \hat{w}) = 1 - \frac{1}{n} \sum_{i=1}^{n} |w_i - \hat{w}_i| \,, \tag{12}$$

where $w \in \{0,1\}^n$ is a ground truth watermark and $\hat{w} = d(y)$ is the watermark extracted from the given, possibly transformed, audio. When a method returns no payload, as occurs for SilentCipher or WavMark, bit accuracy is defined as 0.

4.3 QUALITY LOSSES AND METRICS

Let $s \in [-1,1]^m$ denote the originally generated audio waveform and $\hat{s} \in [-1,1]^m$ the corresponding watermarked waveform. Mean absolute and squared losses over amplitudes and mel-spectrograms, exponential total-variation (ETV) loss (Karimov et al., 2025) over amplitude, and their combinations were used for the quality regularization loss $\mathcal{L}_{\mathrm{qual}}$. However, in our experiments, the best trade-off between quality and imperceptibility was obtained using stand-alone ETV. And the in reported experiments, ETV was utilized. Let $r_i = \hat{s}_i - s_i$ be the difference between the watermarked and clean audio waveforms at sample i. Then

$$\mathcal{L}_{\mathrm{qual}}(s, \hat{s}) = \frac{1}{m-1} \sum_{i=1}^{m-1} \Omega(r_i, r_{i+1}), \tag{13}$$

where m is the number of audio samples and

$$\Omega(x,y) = \begin{cases} \exp\left(|y| - |x|\right) - 1, & \text{if } |y| > |x|, \\ \exp\left(|y|\right) - 1, & \text{if } \mathrm{sgn}(y) \neq \mathrm{sgn}(x), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

We report signal-to-noise ratio (SNR), scale-invariant signal-to-noise ratio (SI-SNR), log-spectral distance (LSD), and two speech-oriented measures used in audio-watermarking evaluations: short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ). Details are provided in Appendix C.

4.4 CONSIDERED AUDIO PERTURBATIONS

We primarily considered the same common transformations (attacks) and their parameters as in the AudioSeal (San Roman et al., 2024) work. During latent optimization, attack parameters θ are sampled from specified ranges, whereas evaluation uses fixed parameters, typically stronger or outside the training range. The attacks and their embedding and evaluation parameters are summarized in Table 7 in the appendix. For example, for additive white noise, the standard deviation is sampled from $[0.001, 0.035]$ during training and fixed to 0.05 during evaluation. For the bandpass attack, during latent optimization, the lower cutoff frequency is sampled uniformly from $[400, 600]$ Hz and the upper cutoff frequency from $[4000, 6000]$ Hz, while evaluation uses fixed cutoffs of 500 Hz and 5000 Hz. For the echo attack, which adds a delayed copy of the signal, the volume and delay parameters are sampled from $[0.1, 0.5]$ during training and fixed to 0.5 during evaluation.

Composite attacks are used only during evaluation, with the fixed evaluation parameters of each component transformation. They are constructed by applying the corresponding single transformations sequentially, using the fixed evaluation parameters. Regular attacks were implemented using the Julius Python library (Défossez, 2022), following AudioSeal. Furthermore, we evaluate watermark robustness under additional codec-based distortions by encoding the watermarked audio with a neural audio codec EnCodec (Défossez et al., 2022), MP3, and AAC, and then decoding it back to the waveform domain.

During latent optimization, the transformations in Table 7 are used. Codec attacks were not applied for latent optimization. Each audio sample is optimized once using the embedding transformations and is then evaluated under the listed attacks.

Table 1: Audio quality of watermarked audio.

Model	SNR↑	SI-SNR↑	STOI↑	PESQ↑	LSD↓
RRS-AW	20.23	20.17	0.9458	4.0433	1.97
AudioSeal	25.76	25.77	0.9890	4.3560	1.83
SilentCipher	**49.19**	**49.20**	**0.9980**	**4.6090**	**1.21**
WavMark	37.62	38.57	0.9920	4.4230	3.21
DualMark	2.18	-2.19	0.7680	1.8942	7.68
TraceMark	4.53	2.89	0.8926	2.9894	4.42
Timbre	26.11	26.15	0.9740	4.2000	4.53

4.5 COMPETITORS

We compare RRS-AW against a suite of neural audio watermarking methods. These include post-hoc approaches: AudioSeal (San Roman et al., 2024), WavMark (Chen et al., 2023), SilentCipher (Singh et al., 2024), and TimbreWatermark (Liu et al., 2024a). Several in-generation methods were also evaluated, including DualMark (Yang et al., 2025), TraceMark (Luo et al., 2026), LAW (Zhu et al., 2026), Anchor Watermark (Rong & Hu, 2026), and GROOT (Liu et al., 2024c).

4.6 HYPERPARAMETERS OF THE METHOD

The following hyperparameters were used to adjust latent representations using our method. In the embedding process, we use 1000 transformations in total, distributed uniformly across all attack types considered during this stage. The main extraction runs also use $k = 1000$. The additional generator experiments use $k = 100$ for both embedding and extraction and are reported separately. All attack parameters are sampled randomly for each transformation, except for the echo attack. Since echo simulation is computationally more expensive, its parameters are shared within groups of 32 attack samples, while different groups use different echo parameters. We use RedDimNet2 (Yakovlev & Okhotnikov, 2026) as the audio encoder g. This model was originally developed for voice biometrics and was selected for its lightweight, computationally efficient nature. It produces 192-dimensional embeddings that are used to construct a robust audio representation. The encoder batch size is set to 128. Only the aggregated representation is ℓ_2-normalized, as specified above, and the individual 192-dimensional encoder outputs are not normalized.

Each generated audio sample has a duration of 3 seconds and a sampling rate of 16 kHz. Audio generation is performed with 50 inference steps. For all methods, the watermark length is set to 16 bits, except for SilentCipher, for which we use the default 24-bit decoded payload. The margin parameter in the watermark loss is set to $\delta = 0.225$. The total optimization objective uses weights $\lambda_{\mathrm{wm}} = 80$ and $\lambda_{\mathrm{qual}} = 5000$ for the watermark embedding and quality-preservation losses, respectively. The latent optimization is run for at most $R = 700$ steps using an initial learning rate of 0.01. We use a plateau-based learning-rate scheduler with patience of 20 steps, decay factor 0.5, improvement threshold 10^{-4}, minimum learning rate 10^{-5}, and at most 3 learning-rate decays. Experiments were conducted using 3 Nvidia H100-80 GB GPUs.

5 RESULTS AND DISCUSSION

Table 1 reports the quality of the watermarked audio without attacks. RRS-AW preserves high perceptual quality, with an SNR of 20.23 dB, SI-SNR of 20.17 dB, STOI of 0.9458, PESQ of 4.0433, and LSD of 1.97 dB. The human perceptual evaluation (Table 15 in the Appendix) further confirms the high perceptual quality of the generated audio. SilentCipher achieves the strongest objective quality metrics among the evaluated baselines.

Table 2 reports bit accuracy (Eq. 12) under regular non-composite conditions. RRS-AW achieves the best or tied-best result in 6 out of 12 evaluated conditions: Identity, Duck, Lowpass, Smooth, Resample, and Speed. It also remains highly competitive under Pink noise and Bandpass filtering, reaching bit accuracies of 0.990 and 0.979, respectively. Overall, RRS-AW demonstrates consistently strong robustness across a diverse set of signal-level transformations.

The results for composite attacks are shown in Table 3. RRS-AW achieves the highest bit accuracy on 9 out of 13 evaluated composite transformations. In particular, it remains more robust to combinations involving filtering, smoothing, speed modification, and amplitude transformations. For

Table 2: Bit accuracy under regular non-composite attacks. The best result in each row is shown in bold.

Attack Type	RRS-AW	AudioSeal	SilentCipher	WavMark	DualMark	TraceMark	Timbre
Identity	**1.000**	**1.000**	0.960	0.978	0.647	0.816	0.996
Boost	0.865	0.941	0.392	0.241	0.627	0.682	**0.954**
Duck	**0.999**	0.893	0.960	0.786	0.653	0.812	0.952
White	0.669	**0.719**	0.188	0.000	0.566	0.573	0.700
Pink	0.990	0.983	0.689	0.970	0.650	0.761	**0.992**
Lowpass	**0.584**	0.483	0.178	0.000	0.577	0.545	0.492
Highpass	0.934	0.496	0.315	0.976	0.641	0.728	**0.978**
Bandpass	0.979	**1.000**	0.436	0.958	0.520	0.706	0.970
Smooth	**0.978**	0.534	0.237	0.000	0.640	0.686	0.742
Resample	**1.000**	**1.000**	0.960	0.973	0.570	0.815	0.988
Speed	**0.614**	0.489	0.284	0.000	0.420	0.502	0.496
Echo	0.890	0.929	0.408	0.734	0.624	0.683	**0.966**

Table 3: Bit accuracy under composite attacks. The best result in each row is shown in bold.

Attack Type	RRS-AW	AudioSeal	SilentCipher	WavMark	DualMark	TraceMark	Timbre
Lowpass + Duck	**0.569**	0.489	0.178	0.000	0.556	0.542	0.506
Highpass + Smooth	**0.865**	0.499	0.332	0.000	0.632	0.644	0.754
Bandpass + Smooth	**0.914**	0.479	0.456	0.000	0.530	0.631	0.758
Bandpass + Duck	**0.978**	0.888	0.436	0.689	0.521	0.704	0.928
Speed + Echo	**0.593**	0.500	0.229	0.000	0.409	0.498	0.488
Speed + Boost	**0.583**	0.510	0.209	0.000	0.428	0.496	0.504
Boost + Smooth	**0.839**	0.661	0.207	0.000	0.640	0.624	0.676
White + Lowpass	0.480	0.476	0.185	0.000	**0.537**	0.534	0.458
White + Highpass	0.536	0.495	0.190	0.000	0.506	0.559	**0.570**
Pink + Duck	**0.986**	0.874	0.733	0.730	0.653	0.764	0.942
Pink + Echo	0.853	0.921	0.298	0.663	0.617	0.661	**0.962**
Resample + Duck	**0.999**	0.894	0.960	0.753	0.573	0.806	0.944
Resample + Boost	0.840	0.939	0.333	0.245	0.580	0.677	**0.964**

example, under Bandpass + Smooth, RRS-AW reaches 0.914, compared with 0.479, 0.456, and 0.000 for AudioSeal, SilentCipher, and WavMark, respectively. RRS-AW also achieves 0.978 under Bandpass + Duck and 0.999 under Resample + Duck.

Table 4 reports robustness under codec-based attacks. RRS-AW obtains the best or tied-best result in 9 out of 13 settings. It achieves the highest bit accuracy at all evaluated EnCodec bitrates, increasing from 0.599 at 1.5 kbps to 0.956 at 24 kbps. Under MP3, RRS-AW reaches 1.000 bit accuracy at both 64 and 128 kbps, tying AudioSeal. For AAC, it achieves the highest accuracy at 64 and 128 kbps, with 0.991 in both cases.

Table 4: Bit accuracy under compression attacks. The best result in each row is shown in bold.

Attack Type	RRS-AW	AudioSeal	SilentCipher	WavMark	DualMark	TraceMark	Timbre
EnCodec 1.5 kbps	**0.599**	0.530	0.255	0.000	0.520	0.518	0.450
EnCodec 3 kbps	**0.690**	0.519	0.248	0.000	0.547	0.539	0.472
EnCodec 6 kbps	**0.829**	0.674	0.249	0.000	0.579	0.577	0.494
EnCodec 12 kbps	**0.916**	0.886	0.266	0.000	0.613	0.602	0.600
EnCodec 24 kbps	**0.956**	0.944	0.213	0.000	0.592	0.632	0.638
MP3 16 kbps	0.846	**0.984**	0.295	0.060	0.607	0.639	0.960
MP3 32 kbps	0.993	**0.998**	0.490	0.750	0.601	0.703	0.990
MP3 64 kbps	**1.000**	**1.000**	0.781	0.975	0.601	0.745	0.992
MP3 128 kbps	**1.000**	**1.000**	0.939	0.978	0.604	0.754	0.994
AAC 16 kbps	0.813	0.680	0.347	0.000	0.596	0.491	**0.876**
AAC 32 kbps	0.966	0.676	0.390	0.739	0.650	0.488	**0.986**
AAC 64 kbps	**0.991**	0.715	0.808	0.973	0.660	0.480	0.990
AAC 128 kbps	**0.991**	0.704	0.898	0.976	0.653	0.479	0.990

Table 5 reports the false-positive rate (FPR) on 512 recordings from the Freesound subset of LAION-Audio-630K. RRS-AW produces no false positives, corresponding to an empirical FPR of 0.0000% and matching the lowest observed FPR among the evaluated methods. Table 6 reports the true-positive rate (TPR) under the detection rule retained for each method. RRS-AW and WavMark use a Hamming-distance threshold of $H \leq 1$, AudioSeal additionally applies its native detection gate, SilentCipher uses $H \leq 4$ together with its native detector, and Timbre requires an exact payload match. DualMark and TraceMark are evaluated using their respective method-specific detection rules. The standard group excludes the identity condition, and the all-attacked TPR is averaged over all 37 attacked conditions.

Table 5: False-positive counts on 512 Freesound recordings.

Method	False positives	Empirical FPR (%)
RRS-AW	**0/512**	**0.0000**
AudioSeal	**0/512**	**0.0000**
SilentCipher	1/512	0.1953
WavMark	**0/512**	**0.0000**
DualMark	1/512	0.1953
TraceMark	**0/512**	**0.0000**
Timbre	2/512	0.3906

Table 6: TPR by attack group.

Method	Standard	Composite	EnCodec	MP3	AAC	All attacked
RRS-AW	0.5927	0.4062	**0.3680**	0.8550	**0.7850**	0.5459
AudioSeal	0.4364	0.2154	0.2760	**0.9800**	0.0050	0.3492
SilentCipher	0.3164	0.2138	0.0040	0.5000	0.5050	0.2784
WavMark	0.4891	0.2123	0.0000	0.6800	0.6400	0.3627
Timbre	0.5400	0.3015	0.0040	0.8750	0.7650	0.4443
DualMark	0.0111	0.0030	0.0075	0.0145	0.0186	0.0050
TraceMark	**0.7385**	**0.5610**	0.3580	0.9415	0.0095	**0.5680**

Overall, RRS-AW demonstrates consistent robustness across all evaluated attack families while preserving high perceptual audio quality. It achieves the best or tied-best bit accuracy in 6 of 12 regular conditions, 9 of 13 composite attacks, and 9 of 13 codec settings, while producing no false positives in the evaluated detection benchmark. The particularly strong performance under composite and codec transformations is consistent with the use of robust audio representations averaged over transformed versions of the input.

6 Limitations

The proposed method has several limitations. First, audio quality is evaluated using objective metrics and a short human perceptual study. RRS-AW achieves an SNR of 20.23 dB, STOI of 0.9458, PESQ of 4.0433, and LSD of 1.97 dB, while the perceptual evaluation confirms high audio quality. A larger-scale listening study across more diverse audio content would provide a more comprehensive assessment of imperceptibility.

Second, watermark embedding and extraction are computationally demanding. Extraction requires applying multiple parametric transformations and encoding the resulting augmented signals to estimate the sample-mean robust representation from Eq. equation 4. In our experiments, extraction takes approximately 0.75 seconds per audio sample on a single GPU. Embedding additionally requires optimization with respect to this representation and takes approximately 6 minutes per sample, depending on the plateau-based stopping criterion. The cost can be reduced by using a lightweight encoder g and by applying early stopping once the watermark has been successfully embedded.

7 Conclusion

In this work, we present RRS-AW, an novel in-generation audio watermarking framework that embeds binary messages into robust audio representations. The method keeps the pretrained generator fixed and performs sample-specific latent optimization to encode the watermark. Experiments with several music generators show that RRS-AW is competitive under individual attacks and performs strongly under composite transformations. The approach also achieves high bit accuracy under several codec-based distortions, including low-bitrate EnCodec and AAC compression. These results align with usage of robust audio representation estimated over transformed versions. The main limitations include lower signal-level fidelity relative to the competitors and higher computational cost due to sampling-based representation estimation. Future work will focus on improving audio imperceptibility, reducing the cost of embedding and extraction, and extending the evaluation to additional generators, datasets, and forgery attacks.

AI USE STATEMENT

AI models (such as AudioLDM) were used for audio generation and experiments, as discussed in the main text. AI models and tools (ChatGPT, DeepSeek, Grammarly) were used for text polishing, shortening, and grammar correction.

ETHICS STATEMENT

The intended application is provenance verification of generated audio through embedded watermarks. Deployment should account for both false attribution and missed watermark detections.

Participation in the human annotation was fully voluntary and unpaid. Participants were under no obligation to complete the study and could discontinue their participation at any point. Incomplete responses were retained only for completed and valid evaluation trials. No personally identifying information was requested or collected, and the collected responses were analyzed only in aggregate. Participation was not tied to any compensation, academic credit, or other benefit. No formal ethics board review was conducted for this listening study.

REPRODUCIBILITY STATEMENT

Algorithms 1 and 2 describe embedding and extraction, and the main text gives the principal hyperparameters. Source code is anonymously available.

REFERENCES

Brian Chen and Gregory W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 2001.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

Yen-Shan Chen, Shih-Yu Lai, Ying-Jung Tsou, Yi-Cheng Lin, Bing-Yu Chen, Yun-Nung Chen, Hung yi Lee, and Shang-Tse Chen. Latent-mark: An audio watermark robust to neural resynthesis. parXiv preprint, 2026.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023.

Alexandre Défossez. Julius: Fast pytorch-based dsp for audio and 1d signals, 2022. URL `https://github.com/adefossez/julius`. MIT License.

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

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Keenan Freyberg. Introducing v3, 2024. URL `URL:https://suno.com/blog/v3`.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 3590–3598, 2023.

Elvir Karimov, Alexander Varlamov, Danil Ivanov, Dmitrii Korzh, and Oleg Rogov. Novel loss-enhanced universal adversarial patches for sustainable speaker privacy. In *Proc. Interspeech 2025*, pp. 1513–1517, 2025.

Darko Kirovski and Henrique S. Malvar. Spread-spectrum watermarking of audio signals. *IEEE Transactions on Signal Processing*, 2003.

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

Slavko Kovačević, Elena Nešović, Kosta Pavlović, Petar Nedić, and Igor Djurović. Deepmark benchmark: Redefining audio watermarking robustness. *IEEE Access*, 2026.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.

Yue Li, Weizhi Liu, Kaiqing Lin, Dongdong Lin, and Kassem Kallas. Trinimark: A robust generative speech watermarking method for trinity-level traceability. *arXiv preprint arXiv:2504.20532*, 2025.

Chang Liu, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Dear: A deep-learning-based audio re-recording resilient watermarking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking. In *Network and Distributed System Security Symposium*, 2024a.

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

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024b.

Weizhi Liu, Yue Li, Dongdong Lin, Hui Tian, and Haizhou Li. Groot: Generating robust watermark for diffusion-model-based audio synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3294–3302, 2024c.

Weizhi Liu, Yue Li, and Zhaoxia Yin. Vocbulwark: Towards practical generative speech watermarking via additional-parameter injection. arXiv preprint, 2026.

Yixin Liu, Lie Lu, Jihui Jin, Lichao Sun, and Andrea Fanelli. Xattnmark: Learning robust audio watermarking with cross-attention. In *International Conference on Machine Learning*, pp. 38987–39015. PMLR, 2025.

Wenhao Luo, Zhangyi Shen, Ye Yao, Feng Ding, Guopu Zhu, and Weizhi Meng. Tracemark-ldm: authenticatable watermarking for latent diffusion models via binary-guided rearrangement. *Expert Systems with Applications*, pp. 131831, 2026.

Kosta Pavlović, Lazar Stanarević, Petar Nedić, Elena Nešović, Slavko Kovačević, and Igor Djurović. Aware: Audio watermarking with adversarial resistance to edits. arXiv preprint, 2025.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Xianjin Rong and Donghui Hu. Anchor watermark: Robust attribution for diffusion-based text-to-audio model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32817–32825, 2026.

Euschi Salah, Zermi Narima, Amine Khaldi, and Kafi Med Redouane. Survey of imperceptible and robust digital audio watermarking systems. *Multimedia Tools and Applications*, 84(7):3635–3681, 2025.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning*, pp. 43180–43196. PMLR, 2024.

Farnaz Sedaghati, Yuxi Wang, Zicheng Weng, and Wei Rao. Voxwatermark: A large-scale benchmark for audio watermark detection under perturbations. *arXiv preprint arXiv:2606.15187*, 2026.

Mayank Kumar Singh, Naoya Takahashi, Wei-Hsiang Liao, and Yuki Mitsufuji. Silentcipher: Deep audio watermarking. In *Interspeech*, 2024.

Mitchell D. Swanson, Bin Zhu, Ahmed H. Tewfik, and Laurence Boney. Robust audio watermarking using perceptual masking. *Signal Processing*, 1998.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136, 2011.

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

Yaoxun Xu, Jianwei Yu, Hangting Chen, Zhiyong Wu, Xixin Wu, Dong Yu, Rongzhi Gu, and Yi Luo. Wake: Watermarking audio with key enrichment. arXiv preprint, 2025.

Ivan Yakovlev and Anton Okhotnikov. Redimnet2: Scaling speaker verification via time-pooled dimension reshaping. *arXiv preprint arXiv:2603.11841*, 2026.

Xuefeng Yang, Jian Guan, Feiyang Xiao, Congyi Fan, Haohe Liu, Qiaoxi Zhu, Dongli Xu, and Youtian Lin. Dualmark: Identifying model and training data origins in generated audio. *arXiv preprint arXiv:2508.15521*, 2025.

Lingfeng Yao, Chenpei Huang, Shengyao Wang, Junpei Xue, Hanqing Guo, Jiang Liu, Phone Lin, Tomoaki Ohtsuki, and Miao Pan. Yours or mine? overwriting attacks against neural audio watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27756–27763, 2026.

In-Kwon Yeo and Hyoung Joong Kim. Modified patchwork algorithm: A novel audio watermarking scheme. *IEEE Transactions on Speech and Audio Processing*, 2003.

Junzuo Zhou, Jiangyan Yi, Yong Ren, Jianhua Tao, Tao Wang, and Chu Yuan Zhang. Wmcodec: End-to-end neural speech codec with deep watermarking for authenticity verification. arXiv preprint, 2024a.

Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, and Zhengqi Wen. Traceablespeech: Towards proactively traceable text-to-speech with watermarking. In *Interspeech 2024*, 2024b.

Ruizhe Zhu, Lingyong Zou, Shuhong Lai, Zhangyi Shen, Zhuang Li, and Tong Qiao. Law: A training-free latent watermark for diffusion-generated audio attribution. *IEEE Signal Processing Letters*, 2026.

A PROOFS

A.1 PROOF OF THE LEMMA 1

Lemma. *Let $\theta_j \sim \mathcal{U}(\Theta)$ be i.i.d. random variables that represent parameters of parametric perturbation ϕ. Let $X_j(t) = g(\phi(s, \theta_j^{enc})) - g(\phi(\phi(s,t), \theta_j^{dec})) \in \mathbb{R}^d$ and let $m(t) = \mathbb{E}(X_j(t))$. Let $Y_j(t) = X_j(t) - m(t)$. Assume that $\|Y_j(t)\|_2 \leq \mu$ and $\mathbb{E}\|Y_j(t)\|_p^2 \leq \sigma^2$.*

Let $\theta \mapsto g[\phi(s,\theta)]$ be a measurable function. Let the function $g \circ \phi(\cdot, \theta)$ be $L-$Lipschitz and let $\|\phi(s,t) - \phi(s,t_0)\| \leq G\|t - t_0\|_q$, where $\phi(s,t_0) = s$ for all s.

Then, with the probability at least $1 - \eta$,

$$\|\Delta(t)\|_2 \leq LG\|t - t_0\|_q + \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}} \tag{15}$$

and $\|\cdot\|_q$ is the norm on Θ.

Proof. Note that

$$\Delta(t) = m(t) + \frac{1}{k}\sum_{j=1}^{k} Y_j(t), \text{ and} \tag{16}$$

$$\|\Delta(t)\|_2 \leq \|m(t)\|_2 + \left\| \frac{1}{k}\sum_{j=1}^{k} Y_j(t) \right\|_2. \tag{17}$$

Now,

$$\begin{aligned} &\|m(t)\|_2 = \left\|\mathbb{E}_\theta[g(\phi(s, \theta^{enc}))] - \mathbb{E}_\theta[g(\phi(\phi(s,t), \theta^{dec}))]\right\|_2 \leq \\ &L\|s - \phi(s,t)\| = L\|\phi(s,t_0) - \phi(s,t)\| \leq LG\|t - t_0\|_q. \end{aligned} \tag{18}$$

Recall the vector Bernstein inequality. Specifically, for independent zero-mean random vectors $Z_1, \ldots, Z_n$ such that $\|Z_i\| \leq M$ almost surely and $\sum_{i=1}^n \mathbb{E}\|Z_i\|_2 \leq V^2$ and every $u \geq 0$,

$$\mathbb{P}(\|\sum_{i=1}^{n} Z_i\| \geq u) \leq 2\exp\left(-\frac{u^2}{2V^2 + \frac{2}{3}Mu}\right). \tag{19}$$

Thus, for $n = k$, $M = \mu$ and proxy variance σ^2 from the lemma statement, for every $\varepsilon > 0$,

$$\mathbb{P}(\|\Delta(t) - m(t)\|_2 \geq \varepsilon) \leq 2\exp\left(-\frac{k\varepsilon^2}{2\sigma^2 + \frac{2}{3}\mu\varepsilon}\right). \tag{20}$$

By choosing ε so that

$$2\exp\left(-\frac{k\varepsilon^2}{2\sigma^2 + \frac{2}{3}\mu\varepsilon}\right) \leq \eta, \tag{21}$$

one can get

$$\varepsilon = \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}}. \tag{22}$$

Therefore, with probability at least $1 - \eta$,

$$\|\Delta(t) - m(t)\|_2 \leq \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}}. \tag{23}$$

Using the triangle inequality

$$\|\Delta(t)\|_2 \leq \|\Delta(t) - m(t)\|_2 + \|m(t)\|_2, \tag{24}$$

with probability at least $1 - \eta$,

$$\|\Delta(t)\|_2 \leq LG\|t - t_0\|_q + \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}}. \tag{25}$$

□

A.2 Proof of the lemma 2

Lemma. *Let the watermark w be embedded in the audio $\hat{\gamma}(s)$ with the use of the secret $\mathbb{I} = \{(l_i, u_i)\}_{i=1}^n$. Assume that the watermark embedding optimization converges such that $\mathcal{L}_{wm} = 0$. Let $\Delta_i = |\hat{\gamma}(s)_{l_i} - \hat{\gamma}(s)_{u_i}|$ be the embedding margin for the $i-$th bit. Then, the number of bit flips N_{err} under a perturbation with the parameter t is bounded:*

$$N_{err} \leq |\{i \in \{1, \ldots, n\} : \Delta_i \leq 2\|\Delta(t)\|_\infty\}|. \tag{26}$$

Specifically, with probability at least $1 - \eta$ no bit flips occur under perturbation with parameter t such that

$$LG\|t - t_0\|_q + \frac{\mu \log \frac{2}{\eta}}{3k} + \sqrt{\frac{\mu^2 \log^2 \frac{2}{\eta}}{9k^2} + \frac{2\sigma^2 \log \frac{2}{\eta}}{k}} < \frac{\delta}{2}, \tag{27}$$

where δ is the margin parameter from Eq. 5.

Proof. Let $\hat{\gamma}_{enc}(\cdot)$ denote the estimate of the empirical robust representation during watermark encoding, and $\hat{\gamma}_{dec}(\cdot)$ denote the one during the extraction of the watermark. We explicitly differentiate between them to underline that they are computed over different parameter samples.

Recall that $\Delta(t) = \hat{\gamma}_{\text{enc}}(s) - \hat{\gamma}_{\text{dec}}(\phi(s,t))$. Note that the $\mathcal{L}_{\text{wm}} = 0$ implies that every summand in

$$\mathcal{L}_{\text{wm}} = \frac{1}{n} \sum_{i=1}^{n} [\delta - (-1)^{w_i}(\hat{\gamma}_{\text{enc}}(s)_{l_i} - \hat{\gamma}_{\text{enc}}(s)_{u_i})]_+ \tag{28}$$

is zero. Hence, for each i,

$$(-1)^{w_i}(\hat{\gamma}_{\text{enc}}(s)_{l_i} - \hat{\gamma}_{\text{enc}}(s)_{u_i}) \geq \delta. \tag{29}$$

By introducing $D_i = \hat{\gamma}_{\text{enc}}(s)_{l_i} - \hat{\gamma}_{\text{enc}}(s)_{u_i}$ one can see that $|D_i| = \Delta_i \geq \delta$.

During the watermark extraction the query audio is (possibly transformed) sample in the form $y = \phi(s,t)$. The decoder computes $\hat{\gamma}_{\text{dec}}(y)$ and applied the decoding rule from Eq. 7. Analogously, define

$$D_i' = \hat{\gamma}_{\text{dec}}(y)_{l_i} - \hat{\gamma}_{\text{dec}}(y)_{u_i}. \tag{30}$$

Then, $\hat{w}_i = 0 \leftrightarrow D_i' \geq 0$.

By definition of $\Delta(t)$,

$$\hat{\gamma}_{\text{dec}}(y)_j = \hat{\gamma}_{\text{enc}}(s)_j - \Delta(t)_j. \tag{31}$$

Note here that $\hat{\gamma}_{\text{dec}}(y)$ and $\hat{\gamma}_{\text{enc}}(s)$ are computed over different realizations of $\theta_1, \ldots, \theta_k$.

Therefore,

$$D_i' = (\hat{\gamma}_{\text{enc}}(s)_{l_i} - \Delta(t)_{l_i}) - (\hat{\gamma}_{\text{enc}}(s)_{u_i} - \Delta(t)_{u_i}) = D_i - (\Delta(t)_{l_i} - \Delta(t)_{u_i}). \tag{32}$$

Bit i is flipped iff $\text{sign}(D_i') \neq \text{sign}(D_i)$

It is noteworthy that a sign change in Eq. 32 requires the perturbation term to be greater than the margin:

$$\left|\Delta(t)_{l_i} - \Delta(t)_{u_i}\right| \geq |D_i| \;=\; \Delta_i.. \tag{$\star\star$}$$

Note that

$$\left|\Delta(t)_{l_i} - \Delta(t)_{u_i}\right| \leq |\Delta(t)_{l_i}| + |\Delta(t)_{u_i}| \leq 2\|\Delta(t)\|_\infty. \tag{33}$$

Therefore, a necessary condition for bit i to flip is $2\|\Delta(t)\|_\infty \geq \Delta_i$. Recall that $\Delta_i \geq \delta$ when $\mathcal{L}_{wm} = 0$. Note that if $\|\Delta(t)\|_\infty < \delta/2$ with probability at least $1 - \eta$, then

$$2\|\Delta(t)\|_\infty < \delta \leq \Delta_i \quad \text{for all i,} \tag{34}$$

and $N_{\text{err}} = 0$ with probability at least $1 - \eta$. Note that the inequality $\|\cdot\|_p \geq \|\cdot\|_\infty$ for $p \geq 1$ allows to use the bound from the Lemma 1 here to bound $\|\Delta(t)\|_\infty$.

□

B Additional results and details

Embedding and evaluation attack ranges. Table 7 summarizes the parameters of the audio perturbations used during watermark embedding and evaluation.

Table 7: Embedding and evaluation parameters for parametric audio perturbations.

Attack Type	Embedding Parameters	Eval Parameters
Identity	N/A	N/A
Boost	2 – 4.5	10
Duck	0.7 – 0.9	0.1
White	0.001 – 0.035	0.05
Pink	0.001 – 0.05	0.1
Lowpass	500 – 3000	500
Highpass	800 – 2000	1500
Bandpass	400 – 600 4000 – 6000	500 5000
Smooth	2 – 10	40
Resample	12800	12800
Speed	N/A	1.25
Echo	0.1 – 0.5 0.1 – 0.5	0.5 0.5

Experiments with additional music generators. To assess the robustness of the evaluated watermarking methods across different generative models, we additionally consider AudioLDM 2 (Liu et al., 2024b), Tango (Ghosal et al., 2023), and Stable Audio Open (Evans et al., 2025). Table 8 reports the mean bit accuracy averaged over all considered non-composite, composite, and codec attacks, excluding the identity transformation. Since evaluating all attacks across multiple generators is computationally expensive, we use $k = 100$ for these experiments, while the default setting of RRS-AW is $k = 1000$.

Despite the reduced number of samples, RRS-AW achieves the highest mean bit accuracy on three of the four evaluated generators, reaching 0.724 on AudioLDM2, 0.801 on Tango, and 0.830 on Stable Audio Open. On AudioLDM1, RRS-AW achieves 0.762, compared with 0.786 for Timbre. We attribute this gap primarily to the reduced $k = 100$ setting used in the cross-generator experiments: under the default $k = 1000$ setting, RRS-AW reaches a mean bit accuracy of 0.841 on AudioLDM1, exceeding Timbre. Timbre, in contrast, shows a substantial performance drop on the other generators, whereas RRS-AW remains consistently robust across different generative models. AudioSeal also remains relatively stable, but is below RRS-AW on AudioLDM2, Tango, and Stable Audio Open, while WavMark and SilentCipher achieve considerably lower mean bit accuracy overall. These results indicate that RRS-AW generalizes well across different audio generation models even under the reduced $k = 100$ evaluation setting.

Additionally, increasing k from 1000 to 2000 yields only an approximately one percentage point improvement in bit accuracy while nearly doubling the embedding time. This indicates diminishing returns for larger k, making $k = 1000$ a reasonable trade-off between robustness and computational cost.

GROOT experiments. Under our experimental setting, which combines AudioLDM-based music generation with varying watermark messages, GROOT demonstrates limited generalization. The

Table 8: Mean attacked bit accuracy by method and generator. RRS-AW uses $k = 100$.

Method	AudioLDM1	AudioLDM2	Tango	Stable Audio Open
RRS-AW	0.762	0.724	0.801	0.830
AudioSeal	0.740	0.711	0.753	0.712
Timbre	0.786	0.592	0.652	0.648
WavMark	0.384	0.361	0.465	0.437
SilentCipher	0.428	0.461	0.384	0.276

Table 9: Bit accuracy of the Anchor Watermark reimplementation under the evaluated attacks.

Regular	BAC	Composite	BAC	Compression	BAC
Identity	0.974	Lowpass + Duck	0.790	EnCodec 1.5 kbps	0.648
Boost	0.886	Highpass + Smooth	0.911	EnCodec 3 kbps	0.695
Duck	0.972	Bandpass + Smooth	0.910	EnCodec 6 kbps	0.767
White	0.750	Bandpass + Duck	0.910	EnCodec 12 kbps	0.830
Pink	0.948	Speed + Echo	0.584	EnCodec 24 kbps	0.856
Lowpass	0.801	Speed + Boost	0.572	MP3 16 kbps	0.865
Highpass	0.945	Boost + Smooth	0.885	MP3 32 kbps	0.949
Bandpass	0.918	White + Lowpass	0.787	MP3 64 kbps	0.967
Smooth	0.966	White + Highpass	0.747	MP3 128 kbps	0.972
Resample	0.953	Pink + Duck	0.947	AAC 16 kbps	0.559
Speed	0.585	Pink + Echo	0.906	AAC 32 kbps	0.526
Echo	0.930	Resample + Duck	0.954	AAC 64 kbps	0.516
		Resample + Boost	0.868	AAC 128 kbps	0.521

empirical FPR ranges from 30% to 50%, while the average bit accuracy remains close to the 50% chance level for all watermarks except the one used in the original GROOT training. Additional adaptation and fine-tuning did not improve these results. Limited generalization of GROOT to varying watermarks has also been reported in (Li et al., 2025).

Anchor Watermark experiments. Since no official implementation of Anchor Watermark was available, we developed an independent paper-based reimplementation. Anchor achieves an untouched BAC of 0.9735. Its mean BAC is 0.8776 over regular attacks, 0.8285 over composite attacks, and 0.7439 over compression attacks, resulting in an overall attacked BAC of 0.8134. The corresponding per-attack results are reported in Table 9. The waveform-quality metrics of the Anchor reimplementation are reported in Table 10. Since the Anchor reimplementation operates in a substantially different generation and audio-quality regime from the perturbative watermarking methods considered in the main comparison, we report its results separately. Anchor nevertheless achieves strong detection performance, with an overall attacked TPR of 0.5843. Its TPR is 0.7364 on standard attacks, 0.6185 on composite attacks, 0.3520 under EnCodec compression, and 0.9300 under MP3 compression, while dropping to 0.0000 under AAC. On the 512-sample Freesound false-positive benchmark, Anchor produces 9 false positives, corresponding to an empirical FPR of 1.7578%.

LAW experiments. Results for LAW are presented in Table 11. LAW achieves high bit accuracy under attacks but this robustness is obtained at the cost of severe audio degradation: the mean untouched SNR is approximately -3.26 dB, with STOI around 0.06 and PESQ around 1.17. This robustness-quality trade-off substantially limits the practical applicability of LAW.

Empirical estimation of Lipschitz constant. To quantify the L from Lemma 1, we numerically analyze the raw, pre-normalization encoder map

$$F_\theta(x) = g(\phi(x, \theta)), \tag{35}$$

using the ℓ_2 norm in both the waveform and embedding spaces. At differentiable points, the corresponding local sensitivity is

$$L_{\text{local}}(x, \theta) = \|J_x F_\theta(x)\|_{2,2} = \sup_{\|v\|_2=1} \|J_x F_\theta(x) v\|_2 \,, \tag{36}$$

where $v \in \mathbb{R}^m$ and $\|\cdot\|_{2,2}$ denotes the induced Euclidean operator norm. We estimate this quantity using matrix-free power iteration with Jacobian-vector and vector-Jacobian products. Attack parameters and stochastic realizations are held fixed within each estimate, and the Jacobian-vector products are validated against finite differences.

Table 10: Audio quality metrics for the Anchor Watermark reimplementation.

SNR↑	SI-SNR↑	STOI↑	PESQ↑	LSD↓
-3.85	-36.33	0.0820	1.2150	15.83

Table 11: LAW robustness under different attacks.

Attack	BitAcc, %	TPR, %
Boost	93.60	96
Duck	99.36	100
Gaussian noise	77.68	78
Pink noise	98.24	100
Low-pass	75.48	90
High-pass	99.26	100
Band-pass	99.06	100
Smooth	99.30	100
Up/down sampling	99.32	100
Speed	49.08	0
Echo	97.34	100

The experiment comprises around 1,200 audio-attack realizations from 60 audio. Over the valid set, the median, 95th percentile, and 99th percentile local sensitivities are 9.78×10^2, 3.95×10^4, and 1.33×10^5, respectively. The largest observed value was re-evaluated using a stronger numerical budget and yielded

$$\widehat{L}_{\mathrm{emp,raw}} = 2.5498 \times 10^5. \tag{37}$$

We additionally measure finite-radius sensitivity using exact-radius waveform perturbations corresponding to different SNR levels. As shown in Table 12, the observed sensitivity decreases as the perturbation radius increases, from 1.32×10^5 at 60 dB to 2.75×10^3 at 20 dB. Together with the large gap between the median and upper tail of the Jacobian distribution, this indicates that the encoder representation has moderate sensitivity at most sampled points but contains comparatively rare regions of much higher local sensitivity. Importantly, these empirical measurements are intended only to illustrate the local behavior and provide a rough indication of the observed range within the sampled operational domain.

Computational cost. Table 13 compares the per-sample runtime of RRS-AW with the post-hoc baselines on the same NVIDIA H100 80GB GPU. RRS-AW has a substantially higher embedding cost because watermark insertion requires iterative optimization of the latent representation. Its embedding takes 358.63 ± 216.61 s per three-second sample (5.98 min on average), compared with milliseconds for the post-hoc methods. The large variance is caused by the adaptive stopping of the optimization procedure: RRS-AW required 110-700 optimization steps (mean 281.08), resulting in embedding times ranging from 140.47 to 874.23 s (2.34-14.57 min).

Extraction is considerably cheaper than embedding for RRS-AW, requiring 0.749 ± 0.102 s per sample, although it remains slower than the baselines because extraction computes the robust representation over 1,000 transformed views. WavMark is the most expensive competitor at extraction time (0.195 s), whereas AudioSeal, SilentCipher, and TimbreWatermarking require only a few milliseconds. Since ordinary 50-step AudioLDM generation takes approximately 1.01 s, the total generation-plus-embedding latency of all post-hoc baselines remains close to one second, while RRS-AW requires 359.63 s on average. This computational overhead is a principal limitation of the proposed in-generation optimization procedure.

Ablation of randomized representation averaging. To isolate the contribution of randomized transformation averaging, we evaluate a *naive* variant of RRS-AW, in which the robust representation is replaced by a single encoder evaluation, $\gamma_{\mathrm{naive}}(x) = g(x)$, during both watermark embedding and extraction. All other components, including the watermark key, pairwise decoding rule, optimization objective, scheduler, and stopping criterion, are kept unchanged.

Table 12: Numerical L estimates.

Quantity	Empirical estimate
Median L_{local}	9.78×10^2
95th percentile L_{local}	3.95×10^4
99th percentile L_{local}	1.33×10^5
Confirmed $\widehat{L}_{\mathrm{emp,raw}}$	2.5498×10^5
Confirmed finite-radius maxima	
60 dB	1.3159×10^5
50 dB	5.1747×10^4
40 dB	1.9723×10^4
30 dB	8.0615×10^3
20 dB	2.7516×10^3

Table 13: Runtime comparison on three-second audio clips. Total time denotes AudioLDM generation plus watermark embedding and does not include extraction. All methods were evaluated on the same NVIDIA H100 80GB GPU.

Method	Embedding (s)	Extraction (s)	Total (s)	Total RTF
RRS-AW	358.628 ± 216.607	0.7491 ± 0.1022	359.633 ± 216.596	119.878
AudioSeal	0.0077 ± 0.0003	0.0039 ± 0.0002	1.013 ± 0.043	0.338
SilentCipher	0.0026 ± 0.0003	0.0028 ± 0.0002	1.008 ± 0.043	0.336
WavMark	0.0416 ± 0.0277	0.1952 ± 0.0004	1.047 ± 0.053	0.349
Timbre	0.0038 ± 0.0001	0.0026 ± 0.0001	1.009 ± 0.043	0.336

Table 14 compares the naive variant with the full RRS-AW results reported in the main experiments. Removing randomized aggregation consistently reduces robustness while preserving perfect watermark recovery under the Identity condition. Averaged over all attacked conditions, bit accuracy decreases from 0.8413 to 0.7449, a drop of 9.6 percentage points. The largest aggregate degradation is observed under EnCodec compression, where mean bit accuracy decreases from 0.7980 to 0.6648. Similar reductions are observed for standard and composite attacks, from 0.8638 to 0.7681 and from 0.7719 to 0.6858, respectively.

The naive variant provides slightly better audio quality under the Identity condition. Its SNR increases from 20.23 to 20.61 dB and SI-SNR from 20.17 to 20.55 dB, while STOI increases from 0.9458 to 0.9707 and PESQ from 4.0433 to 4.2107. These results indicate a robustness-quality trade-off: randomized representation averaging introduces a modest additional distortion but substantially improves watermark stability under signal-processing and compression attacks.

Human perceptual evaluation. We additionally conducted a hidden-reference listening study to evaluate the perceptual quality of the watermarked audio. Participants were presented with up to 40 evaluation trials corresponding to 10 source clips watermarked using RRS-AW, AudioSeal, WavMark, and SilentCipher. Four practice trials were presented before the main experiment. Participants were asked to use headphones and to indicate whether they had any known hearing difficulties.

In each trial, signal A was the original reference audio, while signals B and C consisted of an identical copy of the reference and its watermarked version. The positions of the hidden reference and the watermarked signal were randomized independently across trials, and participants were not informed which watermarking method was used. Participants rated B and C relative to A on a five-point impairment scale, where 5 denotes imperceptible distortion, 4 perceptible but not annoying, 3 slightly annoying, 2 annoying, and 1 very annoying. They additionally indicated which of B or C differed from A, with a "Cannot tell" option.

Table 15 presents the results. 15 participants took part in the annotation. All evaluated methods receive high perceptual-quality scores. RRS-AW obtains a mean rating of 4.48, with 86% of its evaluated samples rated at least 4. The corresponding mean ratings are 4.65 for AudioSeal, 4.71 for SilentCipher, and 4.49 for WavMark. Participants select "Cannot tell" in 54% of evaluated RRS-AW

Table 14: Ablation of randomized representation averaging. "Naive RRS-AW" uses $\gamma_{\text{naive}}(x) = g(x)$ without randomized transformation averaging, whereas "Full RRS-AW" uses the complete proposed representation. Robustness values are mean bit accuracies over the corresponding attack groups. Δ denotes Naive RRS-AW minus Full RRS-AW.

Metric / attack group	Naive RRS-AW	Full RRS-AW	Δ
Bit accuracy			
Identity	1.0000	1.0000	$+0.0000$
Standard	0.7681	0.8638	-0.0957
Composite	0.6858	0.7719	-0.0861
EnCodec	0.6648	0.7980	-0.1332
MP3	0.8850	0.9598	-0.0748
AAC	0.8338	0.9403	-0.1065
All attacked	0.7449	0.8413	-0.0964
Audio quality under Identity			
SNR (dB)	20.61	20.23	$+0.3790$
SI-SNR (dB)	20.55	20.17	$+0.3730$
STOI	0.9707	0.9458	$+0.0248$
PESQ	4.2107	4.0433	$+0.1673$
LSD (dB)	1.72	1.97	-0.2492

trials. These results show that RRS-AW generally introduces little perceptual impairment, although its distortion remains perceptible for some samples.

Table 15: Human perceptual evaluation.

Method	Mean rating↑	Rating ≥ 4↑	Cannot tell ↑
RRS-AW	4.48	86%	54%
AudioSeal	4.65	91%	69%
SilentCipher	4.71	93%	73%
WavMark	4.49	86%	63%

Participation in the annotation was fully voluntary and unpaid. Participants were under no obligation to complete the study and could discontinue their participation at any point. Incomplete responses were retained only for completed and valid evaluation trials. No personally identifying information was requested or collected, and the collected responses were analyzed only in aggregate. Participation was not tied to any compensation, academic credit, or other benefit. No formal ethics-board review was conducted for this listening study.

C Additional metrics

To quantify the perceptual and signal-level distortion introduced by watermarking, we consider several standard full-reference audio comparison metrics. Let $\epsilon > 0$ be a small constant used for numerical stability. The signal-to-noise ratio (SNR) measures the energy of the original signal relative to the watermark-induced error:

$$\mathrm{SNR}(s, \hat{s}) = 10 \log_{10} \left(\frac{[\|s\|_2^2]_\epsilon^+}{[\|s - \hat{s}\|_2^2]_\epsilon^+} \right). \tag{38}$$

The scale-invariant signal-to-noise ratio (SI-SNR) (Le Roux et al., 2019) compares the estimated signal with the reference signal after removing arbitrary scale differences. Unlike conventional SNR, SI-SNR is invariant to global gain changes, making it useful when the watermarked signal may differ slightly in amplitude. Let $\tilde{s} = s - \mathrm{mean}(s)$ and $\tilde{\hat{s}} = \hat{s} - \mathrm{mean}(\hat{s})$. The projection of $\tilde{\hat{s}}$ onto $\tilde{s}$ is $s_{\text{target}} = \frac{\langle \tilde{\hat{s}}, \tilde{s} \rangle}{\|\tilde{s}\|_2^2} \tilde{s}$, and the residual error is $e_{\text{noise}} = \tilde{\hat{s}} - s_{\text{target}}$. The SI-SNR is then defined

as

$$\text{SI-SNR}(s, \hat{s}) = 10 \log_{10} \left(\frac{\left[\|s_{\text{target}}\|_2^2 \right]_\epsilon^+}{\left[\|e_{\text{noise}}\|_2^2 \right]_\epsilon^+} \right). \tag{39}$$

Log-Spectral Distance (LSD) quantifies the spectral distortion introduced by watermarking by comparing the log-power spectra:

$$\text{LSD}(s, \hat{s}) = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left[10 \log_{10} \left(\frac{\left[\|S_s(t,f)\|^2 \right]_\epsilon^+}{\left[\|S_{\hat{s}}(t,f)\|^2 \right]_\epsilon^+} \right) \right]^2}. \tag{40}$$

Here, $S_s(t, f)$ and $S_{\hat{s}}(t, f)$ denote the complex short-time Fourier transform (STFT) coefficients of the original and watermarked audio, respectively, at time frame t and frequency bin f. Therefore, $\|S_s(t, f)\|^2$ and $\|S_{\hat{s}}(t, f)\|^2$ are the corresponding STFT power spectra. The quantity T is the number of time frames, and F is the number of frequency bins. Lower LSD values indicate smaller spectral deviation and better preservation of the original audio quality.

The short-time objective intelligibility (STOI) metric (Taal et al., 2011) estimates the intelligibility of speech by comparing short-time temporal envelopes of the original and watermarked audio. The perceptual evaluation of speech quality (PESQ) metric (Rix et al., 2001) estimates perceived speech quality using a perceptually motivated auditory model. PESQ outputs a quality score related to the mean opinion score (MOS), with higher values indicating better perceived quality. While SNR and SI-SNR measure waveform-level fidelity, STOI and PESQ provide perceptually oriented assessments of intelligibility and speech quality, respectively.