ONE BIT, ONE TIME SLOT: TEMPORALLY STRUCTURED WATERMARKING FOR AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio generation models have made high-quality synthetic speech and music easy to produce, and equally easy to copy and redistribute without attribution, raising serious concerns over accountability and copyright. Watermarking offers a practical solution by embedding recoverable payloads into generated audio for ownership verification. Yet, existing efforts either apply the watermark *post hoc*, leaving it extrinsic to generation and removable by neural denoising or resynthesis, or inject it during *generation* as an *unstructured* global perturbation of the latent that entangles all bits and collapses recovery toward chance as the payload grows. In this work, we present TempoMark, a *temporally structured* generative watermarking framework that treats the generated audio as an ordered trajectory and binds each payload bit to its own temporal footprint. Specifically, TempoMark assigns each bit a dedicated time slot, expands it into chips with a spread-spectrum pattern, and injects the payload as a chip-aligned residual at an intermediate denoising step. Encoding and decoding share this *bit-to-slot* decomposition, under which each bit is recovered by matched filtering over its assigned chips. On top of this structured backbone, a lightweight neural residual in the decoder is trained to support payload readout, together with a small injection adapter as an acoustic-fidelity correction. Extensive experiments show that TempoMark outperforms state-of-the-art baselines in bit recovery accuracy and imperceptibility, and remains robust against audio perturbations and neural remapping attacks.

1 INTRODUCTION

Recent advances in audio generation models (Kong et al., 2021; Lee et al., 2022) have propelled synthetic audio into mainstream content-production pipelines, including speech generation (Liu et al., 2024b), sound effects (Yao et al., 2025; Guo et al., 2026), music clips (Plitsis et al., 2023), and game voices (Yi et al., 2023). As the perceptual gap between machine-generated and human-produced audio rapidly closes, attributing an audio sample to its true origin is becoming a fundamental challenge. Worse still, generated content can be freely reused, redistributed, or maliciously repurposed without any attribution (Sharma et al., 2025; Chen et al., 2023), raising serious concerns over accountability, misinformation, and copyright infringement. These threats call for principled provenance mechanisms that enable reliable traceability while leaving audio quality intact.

Watermarking offers such a mechanism by embedding a recoverable payload into generated content, which can later be extracted to establish ownership. Although extensively studied for text and images (Liu et al., 2024a; Zhao et al., 2025; Hu et al., 2026), these techniques do not transfer to audio, whose signals exhibit distinct perceptual redundancy and representation formats. Existing audio watermarking efforts follow two paradigms, yet both fall short of the provenance goal. *Post-hoc watermarking* (Chen et al., 2023; Roman et al., 2024) embeds the payload after waveform synthesis. Because the watermark is extrinsic to the generative process, it is inherently fragile: neural denoising or resynthesis suffices to strip it away, leaving the content largely unprotected against the attacks it is meant to withstand. *Generative watermarking* instead fuses the payload into the generation process itself, so that the watermark co-emerges with the waveform and is intrinsically harder to excise. This paradigm is conceptually appealing, but its realization remains problematic. The representative method Groot (Liu et al., 2024c) injects message-dependent perturbations into the latent space during denoising, typically through a multi-layer perceptron mapping (bit-to-MLP) that encodes the entire payload as a single *unstructured global perturbation*. We identify this lack of structure as the

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.

core failure mode: all bits are entangled within one holistic latent modification, offering the decoder no separable, bit-level patterns to exploit. After the generator's nonlinear synthesis, only weak and diffuse global cues survive, reducing watermark learning to an ill-posed blind inversion problem, i.e., training becomes ineffective and bit recovery collapses toward chance level.

Our work. We argue that recoverable generative watermarking requires an *explicit structural correspondence* between payload bits and generated audio, and we instantiate this principle with TempoMark, a temporally structured watermarking framework for audio generation. Specifically, TempoMark first expands each bit into spread-spectrum *temporal chips*, converting a fragile single bit into localized and redundant evidence distributed over time. These chips are then injected into temporally aligned regions of the generative state during denoising, so the bit-to-region correspondence is preserved throughout synthesis rather than destroyed by it. Furthermore, a matched hybrid decoder recovers the message by aggregating chip-level evidence from its designated region, mirroring the encoder in a matched-filter fashion. Together, these designs recast watermark recovery from an unreliable global blind-decoding problem into a well-posed *structured evidence aggregation* problem, improving both learnability and robustness.

We summarize our main contributions as follows:

- We identify the absence of explicit structure between payload bits and generated audio as a key failure mode of unstructured generative watermarking, and empirically show that temporal structure is critical for recoverable multi-bit payloads.
- We present TempoMark, which assigns each bit a dedicated time slot and a spread-spectrum chip pattern, and injects a chip-aligned residual during denoising, enabling matched-filter payload recovery in which encoding and decoding share a common temporal structure.
- Extensive experiments on real-world datasets validate that TempoMark recovers multi-bit payloads more reliably than prior baselines, while remaining imperceptible and robust to common audio perturbations and neural remapping attacks.

2 Preliminaries and Motivation

Audio generation models. Audio generation models have demonstrated remarkable capabilities in synthesizing high-fidelity waveforms from a conditioning input, such as text (Kang et al., 2025), Mel spectrograms (Kong et al., 2020; 2021), semantic tokens (Borsos et al., 2023), or acoustic representations (Zeghidour et al., 2022). A typical audio generator G_θ with parameters θ takes a condition c and a latent or noise variable $\mathbf{z}$ as input and produces an audio waveform via:

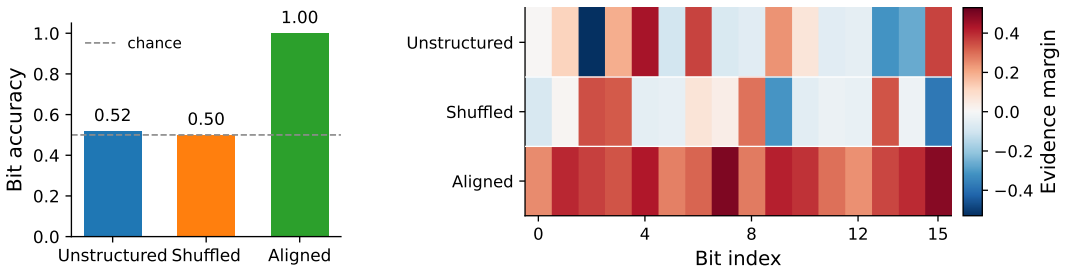
$$\mathbf{x} = G_\theta(c, \mathbf{z}), \qquad \mathbf{x} \in \mathbb{R}^T, \tag{1}$$

where T denotes the waveform length. Depending on the model family, G_θ may be instantiated as a diffusion model, a neural vocoder, an autoregressive model, or a codec-based generator. In diffusion-based audio synthesis, for instance, the generator progressively maps an initial noisy state to a waveform by iteratively denoising along a trajectory conditioned on c.

Despite their utility, the resulting audio can be reused, redistributed, or misused without attribution, highlighting the need for reliable ownership verification of generated waveforms.

Generative audio watermarking. Audio watermarking embeds an imperceptible yet recoverable payload into a waveform for provenance tracing and attribution. According to *when* the payload is embedded, existing methods fall into two categories. *Post-hoc* watermarking modifies the waveform after synthesis and is therefore decoupled from the generation process, making the watermark potentially vulnerable to removal via denoising or resynthesis. In contrast, *generative* watermarking embeds the payload into the generation process, making the watermark an intrinsic property of the generated audio. We focus on the latter in this work.

Formally, an audio generator G_θ synthesizes a waveform $\mathbf{x} = G_\theta(c, \mathbf{z})$ from a condition c and a latent or noise state $\mathbf{z}$. A generative watermarking scheme augments G_θ with two modules. Given an L-bit payload $\mathbf{m} \in \{0,1\}^L$, a watermark encoder E_ϕ produces an injection signal $\mathbf{w} = E_\phi(\mathbf{m})$, which is injected into the synthesis process to yield a watermarked waveform $\mathbf{x}_w = G_\theta(c, \mathbf{z}; \mathbf{w})$. To verify provenance, a decoder D_ψ recovers the payload from a suspect audio as $\hat{\mathbf{m}} = D_\psi(\mathbf{x}_w)$, which is then compared against $\mathbf{m}$.



(a) Bit recovery accuracy.

(b) Per-bit evidence margin.

Figure 1: Structured, aligned evidence is essential for generative audio watermarking. The evidence margin is defined as the response at the correct slot minus the average response at others.

The objective is to recover $\mathbf{m}$ reliably while preserving the perceptual quality of $\mathbf{x}_w$. Under an audio transformation or attack $A(\cdot)$, a robust watermark should remain decodable, i.e., $D_\psi(A(\mathbf{x}_w)) \approx \mathbf{m}$.

Threat model. Following threat models in prior studies (Liu et al., 2024c; Chen et al., 2023), we assume that the attacker obtains a copy of the watermarked audio (e.g., through leakage, scraping, or abuse of a licensed file), but does not know the embedded payload, the decoder, or the diffusion state at injection, and cannot intervene at sampling time. The attacker's goal is to suppress watermark recovery while preserving the utility of the audio, using quality-preserving edits such as additive noise, filtering, resampling, and neural remapping via denoisers.

The legitimate owner fully controls the watermarked inference process $\mathbf{x}_w = G_\theta(c, \mathbf{z}; \mathbf{w})$, embedding the payload before release and keeping both the key and decoder private. Once the waveform is stolen or misused, the owner can recover the payload $\hat{\mathbf{m}} = D_\psi(\cdot)$ from the suspect audio.

Motivation. Existing generative audio watermarking methods (Liu et al., 2024c) typically encode the payload as a global latent perturbation injected during denoising. While this couples the watermark with generation, it does not by itself make the payload learnable. When all bits share one unstructured global perturbation, their evidence becomes entangled within a single latent modification, forcing the decoder to recover individual bits from weak, mutually interfering global cues. Multi-bit recovery thus hinges on a requirement that global perturbation does not meet: the evidence for different bits must remain distinguishable in the generated signal. Localizing bit evidence in structured regions such as temporal slots offers a simple way to achieve this. What remains unclear is whether such structure alone suffices, or whether the correspondence between bits and their assigned slots must also be preserved.

To answer this, we conduct a controlled 16-bit diagnostic study across three regimes: unstructured generative injection, structured evidence with shuffled bit-slot assignment, and aligned structured evidence. As shown in Figure 1a, the first two cases remain at chance-level recovery, with bit accuracies of 0.52 and 0.50, respectively, whereas the aligned design recovers the payload reliably. Figure 1b reveals the underlying cause. Unstructured perturbations provide no stable bit-wise evidence, and shuffled perturbations displace strong responses away from intended slots. By contrast, the aligned structure yields consistently positive margins, indicating that each bit has stronger evidence at its assigned slot than at any other. Together, these findings establish our key insight: generative injection is necessary but insufficient; the recoverable payload must be structured for bit-wise separability and aligned with the decoder's recovery pattern.

3 METHOD

3.1 OVERVIEW

In this paper, we propose TempoMark, a temporally structured watermarking framework for audio generation. Figure 2 depicts the main architecture of TempoMark, which proceeds in three phases.

• *Temporal watermark encoding*: the payload bits are first mapped to chip-level, time-localized spread-spectrum patterns, inducing an explicit temporal layout that is matched at decoding time.

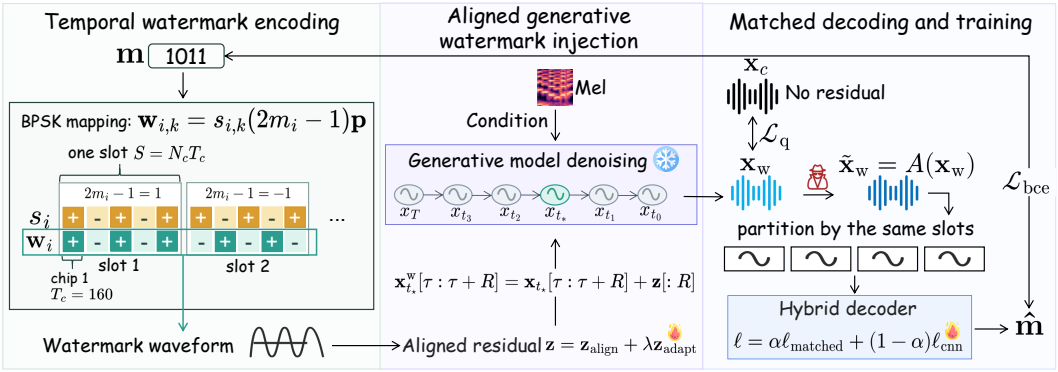


Figure 2: Overview of TempoMark.

• *Aligned generative watermark injection*: these patterns are then lifted into a chip-aligned latent residual and injected into an intermediate diffusion state, allowing the payload structure to be preserved throughout the remaining denoising steps.
• *Matched decoding and training*: the embedded bits are recovered by a hybrid matched-filter decoder. Two auxiliary residual maps are further trained to prevent them from overriding the matched-filter readout.

3.2 TempoMark Design

Temporal watermark encoding. Existing generative audio watermarks encode the payload directly onto the carrier, e.g., global latents or acoustic/transform-domain representations. Global latents embed all bits via a shared denoising perturbation, offering no bit-wise control. Mel or transform embedding specifies where energy is placed, not where each bit is recovered. In both cases, the mapping from a bit to a region of the signal is opaque, making decoding fragile. TempoMark departs from these designs by making *time* the carrier. Two observations motivate this choice. First, time is the one dimension along which a waveform can be partitioned into disjoint supports without any transform, so adjacent bits can be kept physically apart. Second, waveform-based, diffusion-based, and token-based generators unfold the signal along time (van den Oord et al., 2016; Kong et al., 2021; Chen et al., 2021; Borsos et al., 2023; Agostinelli et al., 2023). A time-domain carrier is therefore aligned with how the audio is produced. TempoMark assigns each bit a temporal slot that serves as its local recovery region, reducing payload alignment to a bit-to-slot correspondence.

Given an L-bit payload $\mathbf{m} = (m_1, \cdots, m_L) \in \{0,1\}^L$, TempoMark constructs a watermark waveform $\mathbf{w}$ on a uniform temporal grid. Bit i occupies slot i, and each slot is further split into N_c chips of T_c samples, yielding a slot length of $S = N_c T_c$. Let $\mathbf{p} \in \mathbb{R}^{T_c}$ be a band-limited bipolar pulse that exactly fills one chip. We encode a bit $m_i \in \{0,1\}$ by binary phase-shift keying (BPSK) polarity,

$$\mathbf{t}(m_i) = (2m_i - 1)\mathbf{p}. \tag{2}$$

The two symbols are antipodal and a matched filter reduces to a sign test on $\mathbf{p}$. The bit is then spread across chips by a pseudo-random sequence $\mathbf{s}_i \in \{\pm 1\}^{N_c}$, which repeats this pulse under known signs so that corrupting part of the slot does not erase the bit. For bit $i \in \{1, \cdots, L\}$ and chip $k \in \{0, \cdots, N_c - 1\}$, the encoder writes

$$\mathbf{w}_{[(i-1)S + kT_c : (i-1)S + (k+1)T_c]} \mathrel{+}= s_{i,k}\mathbf{t}(m_i), \tag{3}$$

where $s_{i,k} \in \{\pm 1\}$ is the k-th entry of $\mathbf{s}_i$. The resulting waveform $\mathbf{w}$ is root-mean-square (RMS) normalized and scaled to the target watermark level. This construction establishes an explicit bit-to-slot correspondence that isolates different payload bits along the time axis, while the chips spread each bit's evidence throughout its slot.

Aligned generative watermark injection. The encoded waveform $\mathbf{w}$ should enter the denoising trajectory without disturbing the bit-to-slot layout. Mapping $\mathbf{w}$ through a global MLP would be expressive but can mix temporal positions and erase chip alignment. TempoMark therefore injects a chip-aligned residual. We define $\mathbf{z}_{\mathrm{align}}$ as a band-limited, RMS-scaled copy of $\mathbf{w}$ with the same

number of samples as the diffusion state, copying slot boundaries sample-wise. An auxiliary adapter adds a small correction $\mathbf{z}_{\text{adapt}}$ for local generator dynamics:

$$\mathbf{z} = \mathbf{z}_{\text{align}} + \lambda \mathbf{z}_{\text{adapt}}, \qquad \lambda \ll 1. \tag{4}$$

Since the coefficient λ only scales the adapter and does not prevent cross-chip mixing, we keep it small and regard $\mathbf{z}_{\text{align}}$ as the bit-to-slot evidence. The residual is added on a block of length $R = LS$ at a known offset τ (default $\tau = 0$):

$$\mathbf{x}^{\text{w}}_{t_\star}[\tau : \tau + R] = \mathbf{x}_{t_\star}[\tau : \tau + R] + \mathbf{z}[: R], \tag{5}$$

where $\mathbf{x}_{t_\star}$, $\mathbf{x}^{\text{w}}_{t_\star} \in \mathbb{R}^T$ are the unwatermarked and watermarked diffusion states, and $\mathbf{z}[: R] \in \mathbb{R}^R$ is the leading block of $\mathbf{z}$. The offset τ specifies where this block begins and its particular value matters far less than maintaining encoder–decoder alignment. We add the residual at an intermediate denoising step $t_\star$, then continue sampling to $\mathbf{x}_{\text{w}}$. The unwatermarked counterpart $\mathbf{x}_{\text{c}}$ is generated from the same seed without residual injection.

Matched decoding. Decoding uses the same grid as encoding: slot length S, chip duration T_c, and offset τ. On the samples assigned to bit i, the decoder tests which BPSK polarity better matches the known pulse $\mathbf{p}$. Let $\mathbf{x}_{i,k} \in \mathbb{R}^{T_c}$ denote the RMS-normalized decoder input on chip k of bit i. After despreading with $s_{i,k}$, we correlate against $\mathbf{p}_0 = -\mathbf{p}$ and $\mathbf{p}_1 = +\mathbf{p}$, i.e.,

$$S_i^{(0)} = \sum_{k=0}^{N_c-1} \langle \mathbf{x}_{i,k}, s_{i,k}\mathbf{p}_0 \rangle, \qquad S_i^{(1)} = \sum_{k=0}^{N_c-1} \langle \mathbf{x}_{i,k}, s_{i,k}\mathbf{p}_1 \rangle. \tag{6}$$

The matched-filter logit of bit i is then:

$$\ell_{\text{matched},i} = a_i \left(S_i^{(1)} - S_i^{(0)} \right) + b_i, \tag{7}$$

with learnable calibration parameters a_i, b_i. An auxiliary Mel-conditioned CNN adds a residual $\boldsymbol{\ell}_{\text{cnn}}$ for local acoustic variation that a linear matched filter does not model. So the hybrid logit $\boldsymbol{\ell}$ is

$$\boldsymbol{\ell} = \alpha \boldsymbol{\ell}_{\text{matched}} + (1 - \alpha)\boldsymbol{\ell}_{\text{cnn}}, \tag{8}$$

where $\alpha \in (0, 1)$ is a learned parameter, initialized large so that the matched branch dominates. Bit i is thus read from the same support to which it was written, with no search over unknown lags. TempoMark is therefore evaluated in a time-aligned regime.

Adapter and decoder training. The generator and encoder remain frozen. Only the two auxiliary residual maps introduced above, the injection adapter and the decoder residual, are trained. After denoising to $\mathbf{x}_{\text{w}}$, a stochastic attack layer is applied before decoding. Let $\boldsymbol{\ell}$ be the logits on the attacked $\mathbf{x}_{\text{w}}$. We minimize binary cross-entropy (BCE) loss between $\boldsymbol{\ell}$ and $\mathbf{m}$, which jointly trains both residuals so that injection and readout stay matched under distortion, together with an ℓ_2 term on the adapter only so $\mathbf{x}_{\text{w}}$ does not depart $\mathbf{x}_{\text{c}}$:

$$\mathcal{L}_{\text{bce}} = \text{BCE}(\boldsymbol{\ell}, \mathbf{m}), \qquad \mathcal{L}_{\text{q}} = \tfrac{1}{T}\|\mathbf{x}_{\text{c}} - \mathbf{x}_{\text{w}}\|_2^2. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Datasets and models. We evaluate TempoMark on four English speech corpora and two frozen Mel-conditioned diffusion generators. For the single-speaker scenario, we use LJSpeech (Ito & Johnson, 2017); for the multi-speaker scenario, we use VCTK (Yamagishi et al., 2019); for a large-scale read-speech scenario, we use LibriTTS (Zen et al., 2019) and LibriSpeech (Panayotov et al., 2015). We use DiffWave (Kong et al., 2021), a Gaussian-prior generator, and PriorGrad (Lee et al., 2022), a data-dependent prior, to verify that TempoMark is not tied to one denoising formulation. Corpus statistics and a short backbone description are in Appendix A.1.

Baselines. We compare TempoMark with representative audio watermarking methods: (1) **WavMark** (Chen et al., 2023) is a neural post-hoc method on the synthesized waveform. (2) **AudioSeal** (Roman et al., 2024) is a localized post-hoc method, jointly using an embedder and a detector to insert and localize the payloads on the waveform. (3) **Smark** (Zhang et al., 2025) embeds

Table 1: Comparison of watermarking effectiveness and audio quality between TempoMark and baselines on various datasets and models.

Dataset	Model	DiffWave				PriorGrad			
	Capacity	32		64		32		64	
	Metrics	WSR	PESQ	WSR	PESQ	WSR	PESQ	WSR	PESQ
LJSpeech	WavMark	0.80	2.01	0.78	1.88	1.00	1.86	1.00	1.62
	AudioSeal	1.00	2.32	1.00	1.88	1.00	1.84	1.00	1.64
	MelShield	0.49	1.62	0.51	1.63	0.52	1.91	0.52	1.95
	Smark	0.53	1.61	0.50	1.64	0.51	2.09	0.52	2.12
	Groot	0.51	1.15	0.51	1.15	0.50	1.14	0.50	1.14
	TempoMark	1.00	1.71	1.00	1.77	0.99	3.05	0.98	3.24
LibriTTS	WavMark	0.72	2.20	0.72	1.79	0.98	2.08	0.97	1.86
	AudioSeal	1.00	2.19	1.00	1.84	1.00	2.11	1.00	1.89
	MelShield	0.48	1.43	0.50	1.59	0.48	1.76	0.50	1.77
	Smark	0.53	1.66	0.51	1.66	0.50	1.83	0.51	1.82
	Groot	0.49	1.17	0.49	1.17	0.47	1.09	0.49	1.09
	TempoMark	1.00	1.74	1.00	1.93	0.99	2.80	0.99	2.87
VCTK	WavMark	0.95	1.63	0.95	1.47	1.00	1.21	1.00	1.17
	AudioSeal	1.00	1.64	1.00	1.48	1.00	1.28	1.00	1.18
	MelShield	0.49	1.55	0.50	1.61	0.50	2.45	0.51	2.48
	Smark	0.48	1.71	0.51	1.70	0.51	2.63	0.50	2.50
	Groot	0.49	1.22	0.48	1.22	0.50	1.31	0.50	1.31
	TempoMark	1.00	1.62	1.00	1.67	1.00	2.18	0.99	2.31
LibriSpeech	WavMark	0.84	1.74	0.84	1.48	0.99	1.70	0.99	1.53
	AudioSeal	1.00	1.83	1.00	1.61	1.00	1.65	1.00	1.55
	MelShield	0.49	1.56	0.51	1.60	0.57	1.94	0.51	1.94
	Smark	0.50	1.68	0.50	1.66	0.48	2.02	0.50	2.02
	Groot	0.49	1.19	0.49	1.19	0.48	1.16	0.50	1.16
	TempoMark	1.00	1.75	1.00	1.95	1.00	2.60	0.99	2.96

the payload in a transform-domain acoustic representation. (4) **MelShield** (Jin et al., 2026) embeds the payload into the Mel spectrogram and recovers it after synthesis. (5) **Groot** (Liu et al., 2024c) is a generative audio watermarking method, which injects a message-dependent latent perturbation during denoising, and is the closest to our setting.

Evaluation metrics. We evaluate all methods with the following metrics. Payload extractability is measured by watermark success rate (WSR), the fraction of correctly recovered bits averaged over bits and clips. For imperceptibility, we report peak signal-to-noise ratio (PSNR) as waveform distortion between unwatermarked and watermarked audio, and perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and spectrogram structural similarity index (SSIM) as perceptual scores. Detection reliability is assessed by the ROC AUC and the false-positive rate (FPR) at 95%/99% true-positive rate (TPR). More details are given in Appendix A.2.

Implementation details. All experiments are conducted on a server equipped with three NVIDIA A100-SXM4-80GB GPUs. Unless otherwise specified, TempoMark uses $N_c = 2$, $T_c = 160$ samples and is trained with Adam at learning rate 2×10^{-4}, batch size 4, 20 epochs, and seed 17. Default t_*, τ, and the mix α follow Appendix A.4.

4.2 Experiment Results

Watermark effectiveness. We first evaluate watermark extractability with audio quality. In Table 1, TempoMark keeps WSR above 0.99 across all datasets and generative backbones, with a mean PESQ of 2.26. MelShield and Smark remain near chance in WSR (≈ 0.50), as their bit evidence is entangled on a shared Mel or transform carrier and then reshaped by synthesis. Groot, which injects an unstructured latent perturbation in the same generative setting, likewise remains near chance

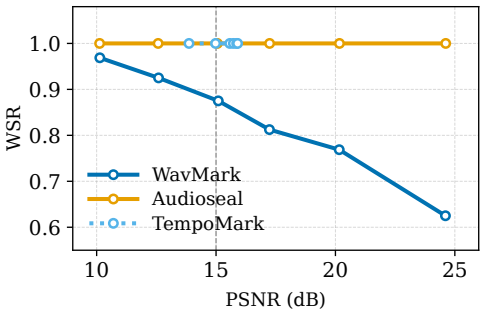


Figure 3: WSR vs PSNR under a shared distortion budget on LJSpeech with DiffWave.

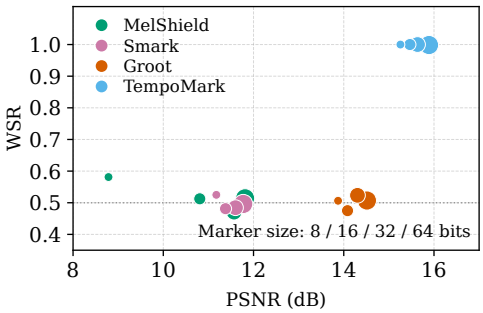
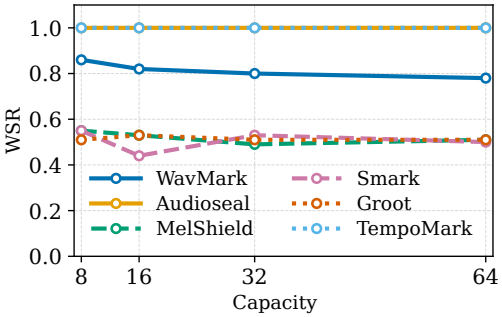
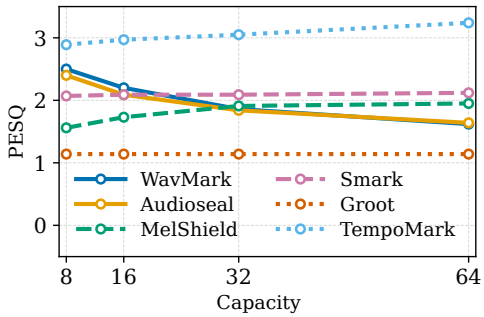


Figure 4: WSR vs PSNR at each method's default setting on LJSpeech with DiffWave.



(a) WSR



(b) PESQ

Figure 5: Trade-off with capacity on LJSpeech (DiffWave). (a) Extractability. (b) Perceptual quality.

with consistently low PESQ. Figure 3 reports WSR against the observed PSNR of the watermarked waveform. TempoMark reaches WSR = 1 in a high-quality band near 14-16 dB. AudioSeal also attains WSR = 1 under milder and harsher PSNR budgets, whereas WavMark becomes reliable only after more of the budget is spent and PSNR has dropped. Figure 4 instead compares with marker size denoting payload length (8/16/32/64 bits). TempoMark stays at WSR = 1 for every payload in a higher-PSNR region than the other defaults, whereas MelShield, Smark, and Groot remain near chance and are therefore left out of the matched-budget sweep. Figure 5 shows that TempoMark's WSR and PESQ remain stable as payload capacity increases, whereas WavMark and AudioSeal retain high WSR, but suffer declining PESQ as the payload grows (Figure 5b), since a longer payload packs more watermark energy into the waveform. Overall, TempoMark sustains extractability without trading off PSNR or payload capacity.

Watermark imperceptibility. We next examine whether reliable extraction comes at the cost of acoustic degradation. Table 2 reports STOI and SSIM for all methods at a 64-bit payload across datasets and generative backbones. On DiffWave, TempoMark achieves mean STOI and SSIM of 0.82 and 0.88, close to the strongest post-hoc methods, whereas Groot stays near 0.5 on both metrics. Switching the backbone to PriorGrad widens this gap in TempoMark's favor: it remains competitive in STOI and clearly surpasses WavMark and AudioSeal in SSIM. VCTK is the hardest set for all methods, yet TempoMark still leads the methods that actually recover the payload (STOI 0.58 against 0.25–0.26 for WavMark and AudioSeal on PriorGrad). MelShield and Smark can post high PriorGrad SSIM, but that spectral similarity is not paired with recoverable bits. A qualitative Mel-spectrogram comparison is provided in Appendix A.3.

False-positive analysis. A desirable provenance watermark should recover the intended payload while refraining from claiming ownership on audio that was never watermarked or was watermarked under a different key. We evaluate both cases on LJSpeech with DiffWave, using the decoder energy $|\text{logit}|$ as a presence score and treating unwatermarked waveforms or wrong-key decodes as negatives. Table 3 reports the detection of unwatermarked versus watermarked at 8, 16, 32, and 64 bits: the AUC is 1.00 in every row, and the false-positive rate at both 95% and 99% true-positive rate is

Table 2: Perceptual quality comparison at 64-bit capacity.

Model	Dataset	LJSpeech		LibriTTS		VCTK		LibriSpeech	
	Metrics	STOI	SSIM	STOI	SSIM	STOI	SSIM	STOI	SSIM
DiffWave	WavMark	0.87	0.92	0.84	0.89	0.66	0.80	0.74	0.86
	AudioSeal	0.89	0.92	0.87	0.90	0.66	0.80	0.80	0.87
	MelShield	0.78	0.92	0.74	0.87	0.63	0.83	0.74	0.87
	Smark	0.77	0.92	0.76	0.90	0.65	0.84	0.75	0.88
	Groot	0.53	0.55	0.54	0.53	0.51	0.36	0.57	0.47
	TempoMark	0.86	0.86	0.83	0.84	0.67	0.81	0.79	0.81
PriorGrad	WavMark	0.89	0.81	0.97	0.81	0.25	0.31	0.91	0.63
	AudioSeal	0.89	0.80	0.97	0.81	0.26	0.33	0.91	0.62
	MelShield	0.82	0.96	0.83	0.97	0.55	0.98	0.85	0.99
	Smark	0.84	0.96	0.84	0.97	0.54	0.98	0.85	0.99
	Groot	0.57	0.65	0.59	0.69	0.47	0.66	0.66	0.75
	TempoMark	0.91	0.95	0.94	0.95	0.58	0.83	0.95	0.95

Bits	AUC	FPR@TPR95	FPR@TPR99
8	1.00	0.00	0.00
16	1.00	0.00	0.00
32	1.00	0.00	0.00
64	1.00	0.00	0.00

Table 3: AUC and FPR on LJSpeech: unwatermarked audio as negatives, watermarked audio as positives.

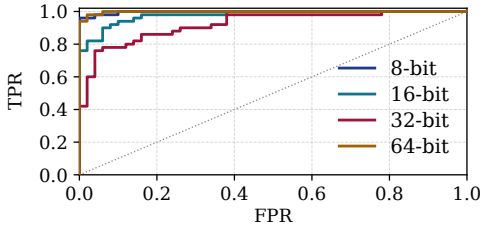


Figure 6: Wrong-key ROC on LJSpeech.

Table 4: WSR (↑) at 64 bits on LJSpeech under signal and neural distortions. GN_{35}: Gaussian noise (35 dB); SS: sample suppression; LP_{5k}: 5 kHz low-pass; MF: median filter; RH: half-rate resample; Vol_{90}: 90% volume; Q_8: 8-bit quantization; Echo: 100 ms echo; TS: time stretch with duration restored to length T; CP_{8k} / CP_{16k}: codec proxy; DN_{20} / DN_{10}: denoise.

Method	Clean	GN_{35}	SS	LP_{5k}	MF	RH	Vol_{90}	Q_8	Echo	TS	CP_{8k}	CP_{16k}	DN_{20}	DN_{10}
WavMark	0.78	0.79	0.79	0.78	0.77	0.78	0.78	0.79	0.75	0.77	0.77	0.78	0.69	0.68
AudioSeal	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.94	1.00	1.00	1.00	0.61	0.62
MelShield	0.51	0.51	0.48	0.49	0.51	0.48	0.52	0.51	0.51	0.50	0.51	0.50	0.53	0.50
Smark	0.50	0.51	0.49	0.50	0.50	0.49	0.51	0.49	0.49	0.50	0.50	0.49	0.51	0.51
Groot	0.51	0.48	0.50	0.50	0.51	0.50	0.50	0.50	0.52	0.49	0.51	0.49	0.50	0.51
TempoMark	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.99

0.00, so clean host audio is never accepted as watermarked. Figure 6 replaces the unwatermarked negatives with a mismatched spreading sequence on the same watermarked clips. The ROC curves remain well above the diagonal at all payload lengths, indicating that a guessed key cannot reproduce the owner's score. The 32-bit setting is the weakest, yet it still separates correct from wrong keys rather than collapsing to chance.

Robustness analysis. Table 4 reports WSR on LJSpeech under clean decoding and multiple distortions. TempoMark stays at or above 0.98 in every column, and reaches 1.00 under most attacks. MelShield, Smark, and Groot, which already fail in the clean setting, remain near 0.50 after distortion and are thus uninformative here. The key contrast lies in post-hoc watermarking. WavMark and AudioSeal survive most classical signal-processing attacks, yet degrade sharply under neural denoising, with AudioSeal dropping to 0.61 / 0.62 on DN_{20} / DN_{10}. This is consistent with reports that post-hoc waveform marks are shallow additive residues and are suppressed by denoisers that reconstruct speech under a clean-speech prior (O'Reilly et al., 2025). By contrast, TempoMark remains at 0.98 / 0.99 on the same attacks.

Table 5: Leave-one-out ablation demonstrating that all components are necessary. Removing any module from the full model degrades WSR. Δ denotes the drop relative to the full model.

Variant	TSS	CHIP	DEC	INJ	WSR ↑	Δ ↓
w/o TSS	×	✓	✓	✓	0.49	0.51
w/o CHIP	✓	×	✓	✓	0.73	0.27
w/o DEC	✓	✓	×	✓	0.50	0.50
w/o INJ	✓	✓	✓	×	0.94	0.06
Full	✓	✓	✓	✓	**1.00**	**0.00**

Ablation study. Table 5 is a leave-one-out ablation of four modules: **TSS**, temporal spread-spectrum payload encoding; **CHIP**, chip-aligned latent construction that preserves bit-to-slot layout; **DEC**, the matched hybrid decoder; and **INJ**, mid-trajectory generative injection. The full model uses all four and attains WSR = 1.00. Δ is the drop after removing one module. Specifically, without **TSS**, WSR falls to 0.49, showing that an unstructured payload is hard to recover after generation. Removing **DEC** causes a similar drop, since local templates can no longer be matched to temporal slots. **CHIP** is supportive but still important, as removing it lowers WSR to 0.73 and indicates that a global latent remap partially erases an otherwise valid temporal code. **INJ** gives the smallest drop, implying that injection timing helps yet remains secondary to the shared TSS/CHIP/DEC layout. Moreover, we report four follow-up ablations in Appendix A.4 including slot-grid alignment, the shared start offset τ, the injection step t_* , and the matched / CNN mix α.

5 Related Work

Post-hoc audio watermarking. Conventional audio watermarking treats embedding as a post-synthesis operation, where a payload is embedded into an existing waveform and later recovered for provenance verification. Recent methods typically learn an encoder-detector pair and train under simulated distortions. WavMark (Chen et al., 2023) recovers multi-bit payloads through neural waveform embeddings. AudioSeal (Roman et al., 2024) supports localized detection for edited or cropped AI-generated speech. Despite their effectiveness, such watermarks remain extrinsic to generation and can be weakened by neural denoising or resynthesis.

Mel- or transform-domain audio watermarking. Another line moves the embedding space from the final waveform to Mel or transform representations. Smark (Zhang et al., 2025) embeds watermarks in wavelet-domain components for diffusion-based TTS. MelShield (Jin et al., 2026) injects watermarks into the Mel spectrogram before waveform reconstruction. Although more structured, these representations remain decoupled from the generator's internal trajectory, and the watermark can be attenuated by reconstruction, regeneration, or representation remapping.

Generative audio watermarking. Generative audio watermarking instead injects payload-dependent signals into the synthesis process, allowing watermark evidence to co-evolve with the audio. A representative method, Groot (Liu et al., 2024c), maps the payload to a latent perturbation and injects it during denoising. While this couples watermarking with generation, it encodes all bits into a single unstructured global perturbation, leaving the decoder without explicit bit-wise correspondence and forcing it to recover the payload from weak, entangled cues.

6 Conclusion

In this work, we present TempoMark, a temporally structured generative audio watermarking framework. TempoMark assigns each payload bit to a local temporal support with chip-level spread-spectrum patterns, transforming multi-bit recovery from entangled global decoding into slot-wise evidence aggregation. Building on this carrier, TempoMark injects a chip-aligned residual into an intermediate state and preserves the intended layout as an inductive bias. Bits are recovered by matched filtering on the same slots and an auxiliary decoder residual supports readout under distortion. Experiments show that TempoMark achieves reliable multi-bit recovery without degrading audio quality, and remains robust to common audio distortions and neural attacks.

7 AI USE STATEMENT

All audio produced by the generative models in this study was inspected by the authors as experimental material, not as a source of scientific claims. TempoMark and the baseline watermarks are the objects of investigation, not tools used to derive results. Experimental findings, method design, and conclusions are the authors' own work and were checked by the authors. Large language models were used only to help edit the manuscript language and the authors reviewed every change and remain responsible for the text.

8 ETHICS STATEMENT

This research is conducted to study intellectual property protection for generative audio and to characterize privacy and security limits of such watermarks. All experiments use publicly available generators and speech benchmarks. We report both successful recovery and the conditions under which extraction fails, with the goal of supporting more reliable attribution of synthetic audio rather than concealing content.

9 REPRODUCIBILITY STATEMENT

For reproducibility, we release the TempoMark implementation for research use at https://anonymous.4open.science/r/One-Bit-One-Time-Slot-Exploiting-the-Generative-Trajectory-for-Effective-Audio-Watermarking-C7EA/. Meanwhile, we have carefully reviewed the documentation of models and datasets to confirm that they do not contain sensitive information or violate protection policies.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *CoRR*, abs/2301.11325, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023.

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

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Xinyue Guo, Xiaoran Yang, Lipan Zhang, Jianxuan Yang, Zhao Wang, and Jian Luan. Av-edit: Multimodal generative sound effect editing via audio-visual semantic joint control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21504–21512, 2026.

Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Mask image watermarking. *Advances in Neural Information Processing Systems*, 38:146313–146346, 2026.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Yutong Jin, Qi Li, Lingshuang Liu, and Jianbing Ni. Melshield: Robust mel-domain audio watermarking for provenance attribution of ai generated synthesized speech. *arXiv preprint arXiv:2605.01515*, 2026.

Hyunju Kang, Geonhee Han, Yoonjae Jeong, and Hogun Park. Audiogenx: explainability on text-to-audio generative models. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i17.33950.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17022–17033. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024a.

Alexander Liu, Matthew Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. Generative pre-training for speech with flow matching. In *International Conference on Learning Representations*, volume 2024, pp. 9435–9454, 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*, MM '24, pp. 3294–3302, New York, NY, USA, 2024c. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680596.

Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech. *arXiv preprint arXiv:2504.10782*, 2025.

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

Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. Investigating personalization methods in text to music generation. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1081–1085, 2023. URL https://api.semanticscholar.org/CorpusID:262067849.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. Proactive detection of voice cloning with localized watermarking. *arXiv preprint arXiv:2401.17264*, 2024.

Tanusree Sharma, Yihao Zhou, and Visar Berisha. Prac3 (privacy, reputation, accountability, consent, credit, compensation): Voice actors in ai data-economy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2360–2372, 2025.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pp. 125, 2016.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/˜idea/readings/rainbow.htm).*, 2019. doi: 10.7488/ds/2645.

Jixun Yao, Hexin Liu, CHEN CHEN, Yuchen Hu, Ensiong Chng, and Lei Xie. Gense: Generative speech enhancement via language models using hierarchical modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 69229–69249, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/acde98fb254b8021d194ccdb80a1241e-Paper-Conference.pdf.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. Add 2023: the second audio deepfake detection challenge, 2023. URL https://arxiv.org/abs/2305.13774.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech, 2019. URL https://arxiv.org/abs/1904.02882.

Yichuan Zhang, Chengxin Li, and Yujie Gu. Smark: A watermark for text-to-speech diffusion models via discrete wavelet transform. *arXiv preprint arXiv:2512.18791*, 2025.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2621–2639. IEEE, 2025.

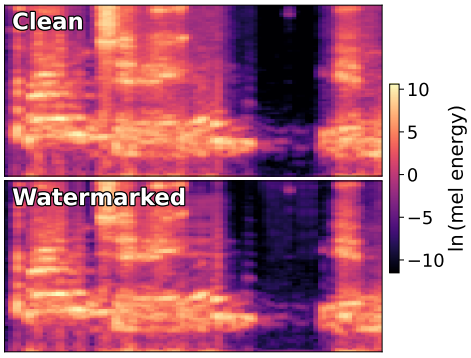


Figure 7: An example of unwatermarked and watermarked Mel spectrograms.

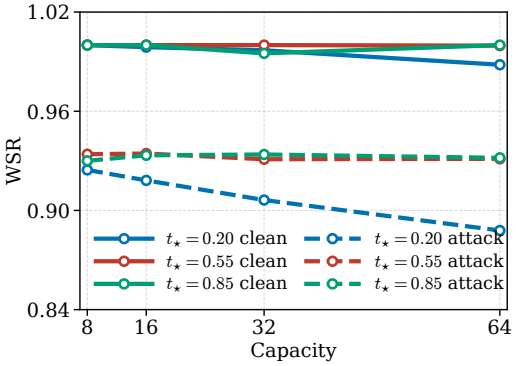


Figure 8: Clean and attack-mean WSR for $t_\star \in \{0.20, 0.55, 0.85\}$.

A Appendix

A.1 Datasets and models

LJSpeech (Ito & Johnson, 2017) is a single-speaker English speech dataset containing 13,100 short audio clips, with a total duration of approximately 24 hours. VCTK (Yamagishi et al., 2019) is a multi-speaker English corpus containing recordings from 110 speakers with diverse accents. LibriTTS (Zen et al., 2019) is a large-scale multi-speaker corpus of 585 hours of 24 kHz, with sentence-level segments and normalized text, which is commonly used for text-to-speech. LibriSpeech (Panayotov et al., 2015) is a large-scale corpus of approximately 1,000 hours of 16 kHz read English speech from public-domain audiobooks. We evaluate TempoMark on two Mel-conditioned diffusion generators to check that the temporal payload layout is not tied to one backbone.

DiffWave (Kong et al., 2021) is a non-autoregressive diffusion generator that synthesizes waveforms from Gaussian noise conditioned on a mel spectrogram. PriorGrad (Lee et al., 2022) is a conditional diffusion generator with a data-dependent adaptive prior, making the denoising process more closely tied to the acoustic condition.

A.2 Metric definitions

WSR is the bit-level recovery rate over a test set of J waveforms,

$$WSR = \frac{1}{JL} \sum_{j=1}^{J} \sum_{i=1}^{L} \mathbf{1}\left[\hat{m}_{j,i} = m_{j,i}\right] \tag{10}$$

where L is the payload length and $\hat{m}_{j,i}$ is the decoded i-th bit of waveform j. WSR $\in [0, 1]$, which is not the rate of fully correct payloads.

PSNR measures waveform distortion between the unwatermarked waveform $\mathbf{x}_c$ and the watermarked waveform $\mathbf{x}_w$:

$$PSNR = 10 \log_{10} \frac{P^2}{\frac{1}{T}\|\mathbf{x}_c - \mathbf{x}_w\|_2^2}, \tag{11}$$

where $P = \max_t |x_c[t]|$ is the peak absolute amplitude of $\mathbf{x}_c$. Higher PSNR means smaller sample-wise error and it is a distortion proxy for imperceptibility, not a listening score.

We report perceptual evaluation of speech quality (PESQ $\in [1, 4.5]$), short-time objective intelligibility (STOI $\in [0, 1]$), and Mel-spectrogram structural similarity index (SSIM $\in [0, 1]$), which score predicted listening quality, envelope intelligibility, and spectral structure between the two Mel maps. For PESQ, STOI, and SSIM, higher is better. We use their standard implementations and do not expand the internal formulae.

Table 6: The impact of temporal alignment.

Condition	8-bit	16-bit	32-bit	64-bit
Clean	1.00	1.00	1.00	1.00
offset 0	1.00	1.00	1.00	1.00
Shift $+10$ ms	0.49	0.45	0.47	0.50
Shift -10 ms	0.52	0.49	0.51	0.50
Shift $+50$ ms	0.49	0.51	0.48	0.51
Shift -50 ms	0.53	0.53	0.49	0.52
Shift $+200$ ms	0.49	0.50	0.49	0.50
Shift -200 ms	0.50	0.51	0.49	0.50

Table 7: Shared start offset τ. $T_c{=}S/2{=}160$ samples. Top: one τ shared by encoder and decoder. Bottom: mismatched pair $(\tau_{\text{enc}}, \tau_{\text{dec}})$.

Condition	8-bit	16-bit	32-bit	64-bit
$\tau{=}0$	1.00	1.00	1.00	1.00
$\tau{=}T_c$	1.00	1.00	1.00	1.00
$\tau{=}50$ ms	1.00	1.00	1.00	1.00
$(\tau_{\text{enc}}, \tau_{\text{dec}}){=}(0, T_c)$	0.50	0.49	0.45	0.48
$(\tau_{\text{enc}}, \tau_{\text{dec}}){=}(T_c, 0)$	0.39	0.43	0.46	0.47
$(\tau_{\text{enc}}, \tau_{\text{dec}}){=}(50\,\text{ms}, 0)$	0.50	0.52	0.53	0.50

A.3 Qualitative Mel spectrograms

Figure 7 compares Mel spectrograms of clean and watermarked audio generated on PriorGrad with a 64-bit TempoMark payload. The two spectrograms remain visually close, reinforcing that the watermark does not introduce an obvious spectrogram artifact at this payload.

A.4 Additional ablations

This appendix supplements the leave-one-out study. We vary the remaining design choices in the order of Section 3: temporal alignment of the slot grid and whether encoder and decoder must share the start offset τ, where in the denoising trajectory the residual is injected (t_*), and how the hybrid decoder mixes matched-filter and CNN logits (α).

Temporal alignment. The diagnosis in Figure 1 isolates structure and alignment under exact registration. Table 6 then shows that this correspondence is strict on a trained checkpoint: on LJSpeech, hybrid WSR is 1.00 whenever encoder and decoder share the start offset, while a shift of only 10 ms drops it to chance at every payload size, and larger lags remain there.

Start offset τ. The injection block begins at τ ($\tau = 0$ in the main experiments). Section 3 argues that the particular value is secondary to encoder–decoder agreement. We therefore freeze a trained checkpoint and vary this offset while keeping weights fixed. When encoder and decoder share the same τ, including $\tau = 0$, a one-shot shift $\tau = T_c$, or a 50 ms delay, hybrid WSR remains high. When they disagree, for example injecting at $\tau_{enc} = T_c$ while the decoder still reads $\tau_{dec} = 0$, recovery collapses, most severely at 8 bits (WSR = 0.39). Thus τ is a shared protocol parameter, not a free per-side knob.

Injection step t_*. Figure 8 sweeps $t_* \in \{0.20, 0.55, 0.85\}$ on LJSpeech as payload length grows. Solid lines are clean WSR and dashed lines are the mean WSR over the attacks in Section 4.2. Mid-trajectory injection $t_* = 0.55$, the default in the main experiments, matches late injection on the attack mean and avoids the weaker clean margin of a very early residual. Earliest injection ($t_*{=}0.20$) is weakest, as remaining reverse steps can overwrite chip polarity in a still-noisy state.

Decoder mix α. The hybrid decoder is $\boldsymbol{\ell} = \alpha \boldsymbol{\ell}_{\text{matched}} + (1-\alpha)\boldsymbol{\ell}_{\text{cnn}}$, with α initialized large so that matched filtering dominates. On a trained DiffWave checkpoint the CNN branch alone stays near

Table 8: The impact of decoder mix α.

Bits	CNN (α=0)	Hybrid ($\hat{\alpha}$)	Matched (α=1)	$\hat{\alpha}$
8	0.48	1.00	1.00	0.81
16	0.49	1.00	1.00	0.81
32	0.51	1.00	1.00	0.81
64	0.50	0.99	0.99	0.82

chance (WSR $\approx$ 0.48-0.51), while the hybrid score tracks the matched branch (WSR $\approx$ 1) with a learned $\hat{\alpha} \approx 0.81$. The CNN term is therefore a small nonlinear residual, not a stand-alone decoder: it does not recover the payload without the matched branch.