KEYBOUND: KEYED AND HOST-BOUND LEARNED AUDIO WATERMARKING FOR SPEECH PROVENANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio watermarking is a proactive route to attributing synthetic speech to its source. Learned audio watermarks are typically judged by payload recovery after a fixed catalog of signal distortions such as noise, compression, filtering, and resampling. That test is necessary but not sufficient for provenance. A mark offered as evidence of origin should not be readable by an unauthorized party, should not be transferable to unrelated audio, and should not vanish when the recording is re-synthesized by a modern generative model. We present KeyBound, a learned audio watermark that restores the two ingredients classical watermarking supplied and learned schemes set aside, a secret key and a host-aware carrier. KeyBound masks the payload with a secret key and embeds the masked bits through a carrier modulated by a host representation derived from a fixed log-mel transform and a learned projection, so the key governs payload access while the host-conditioned carrier resists direct transplantation. A key-independent presence head lets any party detect a mark, whereas only a key holder reads its attribution, and under the single-sample uniformity assumption a wrong-key decode clears our verification rule with probability at most 2.1×10^{-3}. On LibriSpeech against WavMark, AudioSeal, and Timbre, KeyBound holds 1.00 detection accuracy, both at a threshold calibrated on clean audio and at an attack-matched 1% false-positive rate, and 0.98 bit accuracy under a spectral denoiser that costs every baseline its detection, decodes at chance without the key, and rejects transplanted carriers. Detection further transfers to held-out DAC and BigVGAN re-synthesis, though exact payload recovery degrades. Speech provenance is thus better posed as a keyed, host-bound attribution problem than as payload recovery under a fixed catalog of distortions.

1 INTRODUCTION

Advances in text-to-speech and voice conversion have made synthetic speech hard to distinguish from genuine recordings, for human listeners and automatic detectors alike (Wang et al., 2023; Le et al., 2023; Jung et al., 2022). The same technology that serves accessibility and entertainment enables impersonation, financial fraud, and fabricated spoken evidence, making the origin of a recording a question that must stay answerable afterwards. Audio watermarking answers proactively (San Roman et al., 2024; Chen et al., 2023; San Roman et al., 2025), embedding an imperceptible signal at creation so that origin can later be read from the waveform itself.

Most schemes share an encoder–detector design in which a generator adds a low-energy carrier to the host and a jointly trained detector recovers the payload, with recent work reporting strong recovery under noise, filtering, resampling, and compression (Chen et al., 2023; San Roman et al., 2024; Singh et al., 2024; Liu et al., 2025). That evaluation is necessary but insufficient for provenance. A fixed distortion catalog models only accidental degradation, not a party who means to read the payload, erase the mark, or transfer attribution to audio its owner never produced. For a watermark offered as evidence of origin, these adversarial questions are the task itself.

They also expose a limitation of the design rather than of the benchmark. In most schemes no secret protects the carrier, so any party who can run the public detector reads the payload and can turn that access into an estimate of the mark and a means to suppress it. The carrier is also largely independent of the host, which makes transplantation possible, since a pattern recovered from one recording may verify once moved onto another and so fabricate an attribution, one of the forgery

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.

risks benchmarked for audio watermarks (Liu et al., 2024b). Because the carrier is engineered to be quiet, it also tends to live in exactly the signal detail a modern re-synthesis model discards or rewrites, and a single regeneration step erases it (O'Reilly et al., 2025), while a learned remover trained without detector access strips it outright (Li et al., 2025), with provable removal reported in the image domain (Zhao et al., 2024). These three failures share one cause. The mark is bound neither to a secret nor to the signal it marks, precisely the two ingredients classical watermarking supplied and learned schemes set aside, a secret key as in spread-spectrum watermarking (Cox et al., 1997) and a host-aware carrier as in quantization index modulation (Chen & Wornell, 2001) and its roots in dirty-paper coding (Costa, 1983; Gel'fand & Pinsker, 1980; Moulin & O'Sullivan, 2003).

We propose KeyBound, a keyed and host-bound audio watermark that supplies both forms of binding. It retains the learned generator–detector structure of recent watermarks and changes only what the carrier depends on. The payload is first masked with a secret key, so the embedded bits carry no readable attribution without it at all. The carrier is then modulated by host features from a fixed log-mel front end followed by a learned projection, making the mark signal-dependent instead of a pattern that can be lifted and reused. At verification the key recovers the payload, while a key-independent presence head decides on its own whether audio is marked at all. The only keyed learned audio watermark we know of, WAKE (Xu et al., 2025), keys its whole transform and leaves the carrier host-independent, whereas KeyBound keys the payload alone behind a public presence head and anchors the carrier in the host itself.

This design separates three properties that existing evaluations tend to conflate. No-key payload protection comes from the key mask, which randomizes the embedded bits for anyone but the verifier who can strip the mask before decoding. Transplant resistance comes from the host anchor, whose effect we measure empirically against direct transfer of a carrier onto an unrelated host. Robustness to re-synthesis is of a different kind altogether, since no key confers it and it must instead be learned under channels that rewrite low-level acoustic detail. Our experiments show that the distinction is not academic, as highly transparent additive marks survive the standard distortion catalog and still vanish under a single re-synthesis, whereas KeyBound, trained explicitly for that channel, keeps its detection intact throughout our suite of fixed, non-adaptive transformations.

We make three major contributions:

- **Reframing speech provenance.** We argue that provenance demands payload keying and host binding, identify both as absent from current learned audio watermarks, and separate no-key protection, transplant resistance, and re-synthesis robustness into properties each evaluated on their own terms rather than simply assumed together.
- **KeyBound and a verification rule.** A keyed, host-bound learned watermark with a key-independent presence head, together with a provenance-oriented verification rule V_A carrying a conditional single-sample wrong-key acceptance bound (Proposition 2). If the keyed decode is uniform in the adversary's view, a wrong-key decode clears the rule with probability at most $\beta \approx 2 \times 10^{-3}$.
- **Evaluation.** We test KeyBound on LibriSpeech against WavMark, AudioSeal, and Timbre across signal distortion and re-synthesis, and additionally under no-key decoding and carrier transplantation. It reaches a clean-calibrated detection accuracy of 1.00 and 0.98 bit accuracy under a spectral denoiser that costs every baseline its detection, decodes at chance without the key, and rejects transplanted carriers outright.

2 Related Work

Learned audio watermarking. Most learned audio watermarks are post hoc, in that a generator adds a low-energy carrier to the host waveform and a jointly trained detector recovers the payload from it. Examples include invertible window schemes (Chen et al., 2023), localization-trained generator–detector pairs (San Roman et al., 2024), psychoacoustic schemes (Singh et al., 2024), timbre-based schemes (Liu et al., 2024a), dual-embedding invertible networks (Li et al., 2024), and cross-attention detectors (Liu et al., 2025). They advance capacity and imperceptibility and report robustness to a fixed distortion set, yet in nearly all of them the carrier is host-independent and unkeyed, so neither no-key payload protection nor host binding enters the design. A separate line embeds the watermark inside the generative model so that every sample it produces is marked

(San Roman et al., 2025), which helps whoever controls that model but not the verifier who is handed only a waveform, the setting we study.

The principal exception is WAKE (Xu et al., 2025), a keyed learned audio watermark aimed at access control and multi-watermark capacity, which gates an invertible embedding network by the key so that decoding under a wrong key falls to chance. Because the gate governs the whole transform, presence and payload are alike key-dependent. The carrier remains host-independent, the keyed claim is purely empirical, and the evaluation covers classical signal edits alone, testing neither removal nor false attribution. KeyBound masks only the payload behind a key-independent presence head, so any party may detect a mark while only a key holder reads its attribution. It anchors the carrier in host features computed from a fixed log-mel front end, driving transplant verification to near zero, supplies a formal no-key statement and a conditional single-sample wrong-key acceptance bound (Proposition 2), and, trained under neural re-synthesis, holds detection under a spectral denoiser where additive schemes in WAKE's own family collapse (O'Reilly et al., 2025).

Classical foundations. Both properties that KeyBound restores are old watermarking ideas. Carrier secrecy is the premise of spread-spectrum watermarking, whose secret spreading key hides the mark and protects it (Cox et al., 1997). Host dependence is the premise of quantization index modulation, whose codes embed the payload relative to the cover signal with a minimum-distance margin (Chen & Wornell, 2001). Quantization index modulation is itself an instance of communication with side information at the encoder, as in Costa's dirty-paper coding (Costa, 1983) and the Gelfand–Pinsker channel (Gel'fand & Pinsker, 1980), analyzed for information hiding by Moulin & O'Sullivan (2003). Learned audio watermarks traded both away for transparency and distortion-robustness, whereas KeyBound carries keying and host binding into a learned design and trains it under the re-synthesis attacks that the classical codes never had to face.

3 Problem Formulation

We watermark a host waveform $x \in \mathbb{R}^T$ with an n-bit payload $m \in \{0,1\}^n$ under a secret key $\kappa \in \{0,1\}^L$ held only by the embedder and an authorized verifier. An embedder E produces a marked signal $x_w = E(x, m, \kappa)$ within a perceptual budget, and an adversary may apply any transformation a from an admissible class $\mathcal{A}$ within a distortion budget. A key-independent presence head returns a score $s(y) \in [0,1]$, whereas an authorized decoder D, available to a key holder alone, returns a payload estimate $\hat{m}(a(x_w), \kappa)$. Beyond payload recovery under benign a, we ask three things of (E, D). *Robustness* operates at two levels, since presence robustness asks only that the marked/unmarked decision survive an admissible transformation, whereas payload robustness further asks $\hat{m} = m$. Detection accuracy (DA) measures the former and is our primary metric, bit accuracy (BA) the latter. *No-key payload protection* requires that without κ the detector output decode at chance (Proposition 1), a statement about detector-side recovery, not about multi-message secrecy of the waveform. *Resistance to false attribution* asks that a party without κ should not make audio the verifier accepts under a chosen payload, which we test in the case where a carrier is lifted from a marked recording onto unrelated audio. The key governs payload access alone and confers no robustness against removal, which comes from the host-bound carrier and attack-channel training.

Keyless post-hoc schemes leave two gaps for provenance. Absent a secret key, a public detector reads m straight from x_w, and no amount of distortion training alters that. Absent host binding, a carrier lifted from one recording may verify on another whenever it is reusable, as we measure directly on a leading scheme, though not every keyless scheme proves transplantable. A secret key and a host-conditioned carrier address both gaps, and none of the baselines we evaluate supplies the two together, the gap KeyBound is built to close.

3.1 Threat Model

A provider embeds a watermark into the speech it generates under a secret key κ, and a verifier holding κ later tests a recording to establish its origin. We assume the adversary knows everything but the deployment key, so the architecture and weights of E, D, and the host analysis C may be public. It may run them on arbitrary inputs and trial keys and collect any number of marked recordings, yet it can neither obtain κ nor query a verifier that holds it. Our keyed-access claim is thus

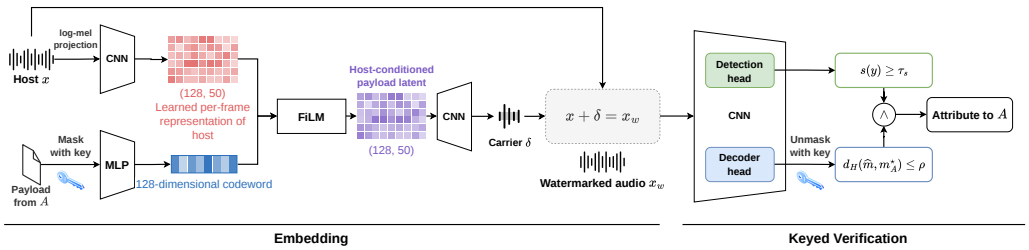

Figure 1: Overview of KeyBound. The payload is masked with a secret key and embedded through a carrier modulated by spectral host features $C(x)$, so the mark is both keyed and host-bound. Here A denotes a registered party with key κ_A and payload $m_A^\star$ (Section 4.3). At verification the detector produces a key-independent presence score $s(y)$, letting any party detect that audio is marked, while only a holder of κ can unmask it. Attribution requires both, since the rule V_A fires when the presence score clears τ_s *and* the decode lands within ρ bits of the registered payload.

single-sample, because the fixed reused mask affords no multi-message secrecy and no resistance to known-payload inference, both of which fall outside what this paper claims and evaluates.

The adversary holds a marked recording and pursues one of three goals. It may *read the payload*, recovering m without κ. It may *remove the mark*, perturbing the recording within a perceptual budget until the detector stops firing, whether by signal processing, by re-synthesis through a codec, a vocoder, or a denoiser, or by a removal network trained against the scheme. It may *produce a false attribution*, presenting a recording the verifier accepts under a chosen payload, either by embedding one or by transplanting a carrier from a marked recording onto audio the owner never produced. The adversary lacks only the key, and we test each of these three goals under the attacks that we implement. An attack counts only while the audio stays perceptually useful, since degrading it further would defeat the adversary's own purpose in attacking it.

4 Method

KeyBound keeps the generator–detector structure of post-hoc watermarks, changing two things only, what the carrier depends on and how the payload is protected. It operates on 16 kHz audio in one-second units with a 16-bit payload, and Figure 1 sketches its embedder and detector.

The key masks the payload so that the masked bits appear random without the key,

$$\tilde{m} = m \oplus \mu(\kappa), \qquad \mu(\kappa) = \kappa_{1:n}, \tag{1}$$

where $\oplus$ is the bitwise XOR and $\mu(\kappa)$ is a per-bit mask read from the key. The host analysis C maps the host to latent-width features through a fixed, phase-free log-mel transform followed by a small learned projection. The generator forms the carrier by projecting the masked payload and modulating it with $C(x)$ through feature-wise linear modulation (FiLM) (Perez et al., 2018),

$$z = \mathrm{FiLM}\big(P(\tilde{m}),\, C(x)\big), \qquad \delta = \alpha\, G_\theta(z), \tag{2}$$

where P is a learned per-bit projection, G_θ is a convolutional synthesis network, and α scales the carrier to the perceptual budget. The marked signal adds the carrier to the host,

$$x_w = x + \delta. \tag{3}$$

Because δ passes through $C(x)$ in (2), the carrier becomes a function of the recording it marks and is harder to reuse, while transplant-negative training teaches the detector to reject a carrier moved onto a mismatched host. The masked payload still drives the carrier, but the host modulation is substantial by design, since a carrier the host barely influenced would be the transferable pattern that the anchor exists to prevent.

4.1 Keyed Decoding and Detection

The detector computes a learned analysis of the received audio y, produces per-frame payload logits and detection logits, and pools the payload logits with attention weights computed from the

detection-head logits,

$$h = E_\psi(y), \quad \bar{\ell} = \sum_t \mathrm{softmax}_t\big(w_d^\top h\big)_t \, \big(W_m h_t\big), \tag{4}$$

where W_m and w_d are the payload and detection heads. It then removes the key mask by flipping the sign of every logit whose bit the mask inverted, an operation that recovers the payload for a key holder and for nobody else,

$$\ell = \bar{\ell} \odot \big(1 - 2\,\mu(\kappa)\big), \qquad \hat{m} = \mathbf{1}[\ell > 0]. \tag{5}$$

The clip-level presence score is obtained separately by averaging the frame-level sigmoid outputs of the detection head,

$$s(y) = \tfrac{1}{T'} \sum_t \sigma(w_d^\top h_t), \tag{6}$$

where T' counts detector frames and s is calibrated once on clean unmarked audio.

4.2 Training Objective

We train the generator G_θ, the detector E_ψ, and the learnable parameters ϕ of the payload projection P and of the host-feature projection in C end to end with the log-mel front end of C frozen. Each step draws an attack channel $a \sim \mathcal{A}$ and applies it to the marked signal before decoding, mixing the distortion catalog with re-synthesis (codec round trip, spectrogram vocoder, spectral denoiser) and with an in-loop removal network R that is updated alongside the model. The objective is

$$\begin{aligned} \min_{\theta,\psi,\phi} \mathbb{E}\big[&\lambda_m \mathcal{L}_{\text{msg}} + \lambda_d \mathcal{L}_{\text{det}} + \lambda_f \mathcal{L}_{\text{forge}} \\ &+ \lambda_p \mathcal{L}_{\text{perc}} + \lambda_r \mathcal{L}_{\text{snr}}\big], \end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{msg}}$ is the binary cross-entropy between the keyed decode $\ell(a(x_w), \kappa)$ and the payload m. The presence-head cross-entropy $\mathcal{L}_{\text{det}}$ takes marked audio, whether attacked or passed through R, as positive against three negative sets drawn from unmarked audio alone, namely clean audio, attacked clean audio, and the output of R on attacked clean audio, which denies the head any chance of learning to equate distortion itself with the presence of a mark. The false-attribution term $\mathcal{L}_{\text{forge}}$ drives the presence score to zero for a carrier transplanted onto a shuffled host, and the final two terms bound perceptibility,

$$\begin{aligned} \mathcal{L}_{\text{perc}} &= \mathcal{L}_{\text{mrstft}} + \mathcal{L}_1 + 2\mathcal{L}_{\text{mel}} + \lambda_{\text{TF}} \mathcal{L}_{\text{TF}}, \\ \mathcal{L}_{\text{snr}} &= \max(0,\, \tau_{\text{snr}} - \mathrm{SNR}), \end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{mrstft}}$ is the multi-resolution spectrogram loss (Yamamoto et al., 2020), the psychoacoustic time–frequency masking loss $\mathcal{L}_{\text{TF}}$ places carrier energy where the host masks it, and τ_{snr} is ramped from 6 to 22 dB. The removal attack we report at test time is not the in-loop remover R but a held-out reconstruction remover R_{test}, a separately trained network with access to neither the detector nor the key. Robustness to these channels is empirical and comes from this training and the host-bound carrier, while the key governs payload access alone.

4.3 Keyed Provenance Verification

We cast provenance as a per-party verification rule. Party A holds a secret key $\kappa_A \in \{0,1\}^L$ known only to A and an authorized verifier, registers a payload $m_A^\star \in \{0,1\}^n$, and produces marked audio $x_w^A = E(x, m_A^\star, \kappa_A)$. On a received waveform y we distinguish two decisions. **Presence verification** is key-free and asks only whether y carries any mark at all,

$$V_{\text{pres}}(y) = \mathbf{1}[s(y) \geq \tau_s], \tag{9}$$

while **attribution verification** is keyed and asks whether y carries A's payload under A's key,

$$V_A(y) = \mathbf{1}[s(y) \geq \tau_s] \wedge \mathbf{1}[d_H(\widehat{m}(y, \kappa_A), m_A^\star) \leq \rho], \tag{10}$$

where $s(y) \in [0,1]$ is the key-independent presence score, $\hat{m}(\cdot, \kappa_A)$ the keyed decoder, d_H Hamming distance, τ_s a presence threshold, and ρ the tolerated bit-error budget, set to $\rho = 2$ so that at least $n - \rho = 14$ of 16 bits must agree. DA measures presence verification, while attribution verification is the operational provenance decision and adds the keyed payload check. We state the keyed-access guarantees of the mask first and then two properties that we verify only empirically.

Proposition 1 (Single-sample mask uncertainty)**.** *For a single sample, if the mask $\mu(\kappa)$ is uniform and unknown to the adversary, then $\tilde{m} = m \oplus \mu(\kappa)$ is uniform in that view, and a decoder without κ or a wrong key recovers m at chance.*

Proposition 2 (Conditional single-sample wrong-key acceptance bound for V_A)**.** *If $\hat{m}(y, \kappa_A)$ is uniform in the adversary's view, as Proposition 1 gives for a single sample under an unknown uniform mask, then*

$$\Pr\left[V_A(y) = 1\right] \leq \beta := \sum_{i=0}^{\rho} \binom{n}{i} 2^{-n}. \tag{11}$$

For $n = 16, \rho = 2$, $\beta = 137/2^{16} \approx 2.1 \times 10^{-3}$.

Appendix A.1 proves both propositions. The bound rests on the keyed payload check rather than on τ_s, so inflating the presence-only false-positive rate gains nothing while 14 of A's 16 keyed bits must still be reproduced. The two properties stated next are empirical.

Robust completeness. A's own audio should verify under every admissible $a \in \mathcal{A}$, that is $\Pr[V_A(a(x_w^A)) = 1] \geq 1 - \varepsilon_a$. Because V_A demands presence ($s \geq \tau_s$) and payload ($d_H \leq \rho$) together, its true-accept rate tracks whichever conjunct is weaker, and the two diverge whenever re-synthesis preserves presence while destroying exact payload. Table 2 reports the rate per attack, with DA standing for the presence conjunct and the bit budget standing for the payload one.

Transplant resistance. A's genuine carrier should not transfer, so for $\delta_A = x_w^A - x$ lifted onto an unrelated host x' we want $\Pr[V_A(x' + \delta_A) = 1] \approx 0$. Proposition 2 does not imply this, since δ_A is a legitimate carrier rather than a guess, and the property follows empirically from the host-conditioned carrier and the $\mathcal{L}_{\text{forge}}$ term, which we test against direct residual transplantation (Appendix A.6).

Scope. The tolerance ρ trades true acceptance against wrong-key acceptance, and Appendix A.13 relates β to payload length. Proposition 2 is single-sample and assumes an honest verifier, so it delivers keyed access rather than multi-message secrecy and bounds no adversary who adaptively queries the verifier to infer the mask. Reused-key deployments, which reuse one 16-bit mask across recordings, fall outside the bound as stated. We do not claim cryptographic unforgeability, since an adversary that observes a valid masked codeword and can invoke or approximate the embedder may attempt replay or re-embedding on a new host. Preventing that needs the payload authenticated jointly with a host fingerprint, by a MAC or a signature, which we leave to future work.

5 Experiments

We evaluate KeyBound along four axes, robustness to attack, imperceptibility, no-key decoding, and transplant-based false attribution, taking the last two as what distinguishes it.

5.1 Experiment Setup

Datasets. We train on the 28,539 one-second clips at 16 kHz of the LibriSpeech (Panayotov et al., 2015) train-clean-100 split and test on 300 held-out clips from test-clean. Clip selection and payload draws are seeded, so all methods see the same 300 clips in the same order under the same payloads. For the reported clip-level evaluation, we independently sample a key per clip from the global random state, and the model is trained over random keys. This protocol evaluates single-sample key uncertainty and does not model a deployment that reuses one key across recordings. LibriSpeech is the corpus on which the released baselines were themselves developed and evaluated, so no method gains an advantage from a domain shift, and its speaker-disjoint split lets us hold out speakers as well as utterances. It is read English speech throughout, which bounds how far the conclusions drawn below can reasonably be carried beyond that setting. Appendix A.2 gives the architecture, training schedule, loss weights, and computing platform.

Baselines. We compare KeyBound against three representative released schemes, each run from the authors' own code and checkpoints, namely the invertible window scheme WavMark (Chen et al., 2023), the localization-trained AudioSeal (San Roman et al., 2024), and the timbre watermark

Attack	WavMark		AudioSeal		Timbre		KeyBound	
	DA	BA	DA	BA	DA	BA	DA	BA
Clean	1.000	1.000	0.990	0.912	1.000	**1.000**	**1.000**	0.999
Signal distortion								
Noise (20 dB)	0.007	0.501	0.490	0.639	0.363	0.909	**1.000**	**0.999**
Noise (10 dB)	0.000	0.498	0.287	0.545	0.003	0.667	**1.000**	**0.974**
Lowpass (4 kHz)	1.000	**1.000**	0.993	0.912	1.000	0.991	**1.000**	0.999
Highpass (500 Hz)	1.000	1.000	0.993	0.918	1.000	**1.000**	**1.000**	0.999
Resample (0.8)	0.947	0.970	0.967	0.898	1.000	**1.000**	**1.000**	0.999
Gain ($\times 0.5$)	0.993	0.996	0.947	0.908	1.000	**1.000**	**1.000**	0.999
Quantize (8-bit)	0.747	0.864	0.937	0.852	0.977	0.999	**1.000**	**0.999**
MP3 (32 kbps)	0.723	0.849	0.907	0.861	1.000	**1.000**	**1.000**	0.999
Smoothing	0.940	0.969	0.970	0.907	1.000	0.988	**1.000**	**0.999**
Echo	0.677	0.821	0.977	0.870	1.000	**1.000**	**1.000**	0.994
Re-synthesis								
EnCodec (6 kbps)	0.000	0.498	0.850	0.568	0.013	0.562	**1.000**	**0.982**
Griffin-Lim	0.983	0.988	0.080	0.524	1.000	**1.000**	**1.000**	0.702
Denoiser	0.000	0.498	0.173	0.545	0.007	0.712	**1.000**	**0.983**

Table 1: Robustness across the attack suite, with the best DA and BA of each row in bold.

(Liu et al., 2024a) built to survive voice cloning. Timbre runs at its native 22.05 kHz with a 10-bit payload and is evaluated under the same attack algorithms at that rate, while the remaining methods and KeyBound run at 16 kHz. WavMark runs on 1.1 second clips and every other method on 1.0 second clips. Payload length, clip length, and sampling rate therefore differ across methods, so the comparison is not fully controlled, and we keep each baseline at its native setting (Appendix A.3).

Attacks. We group the transformations into two families. *Signal distortions* comprise additive noise, low- and high-pass filtering, resampling, gain, 8-bit quantization, and MP3 at 32 kbps, the catalog the additive schemes are tuned to survive. *Re-synthesis* regenerates the waveform through an EnCodec round trip at 6 kbps (Défossez et al., 2023), a Griffin-Lim spectrogram vocoder (Griffin & Lim, 1984), and a spectral-gating denoiser, hereafter the denoiser, with the parameters and strengths of every attack listed in Appendix A.8.

Metrics. Detection accuracy (DA) (Song et al., 2026) denotes the true-positive rate at a threshold calibrated to 1% FPR on clean unmarked audio, so its realized false-positive rate may change under attack, and we reserve TPR@1%FPR for the attack-matched evaluation of Table 4. Bit accuracy (BA) is the fraction of payload bits recovered correctly. Imperceptibility is read from wideband PESQ (Rix et al., 2001), short-time intelligibility STOI (Taal et al., 2011), and signal-to-noise ratio (SNR), for all of which higher is better. No-key decoding is read from the bit accuracy an adversary obtains by running the detector without the key, where 0.5 is chance, and transplant-based false attribution from the rate at which a carrier lifted onto unrelated audio still verifies.

5.2 Main Results

Robustness. KeyBound holds DA = 1.00 on every perturbation, and under the EnCodec round trip it recovers the payload itself (BA = 0.98) rather than presence alone, departing from earlier additive designs that lose the message to codec re-synthesis. Only Griffin-Lim reduces exact recovery (BA = 0.70), where detection still stays at 1.00, so the mark is confirmed present on audio whose payload can no longer be attributed. Under the spectral denoiser KeyBound keeps DA = 1.00 and BA = 0.98 where every baseline loses detection, WavMark falling to 0.00, AudioSeal to 0.17, and Timbre to 0.01. The baselines fail unevenly across codec and vocoder, each surviving one re-synthesis and collapsing on the other, whereas KeyBound retains presence detection under both, although its payload recovery under Griffin-Lim drops to BA = 0.70 (V_A 0.20, Table 2). Because DA fixes its threshold on clean unmarked audio, we check that the attacks do not inflate it. Including the clean condition, seven of the fourteen evaluation conditions realize a false-positive rate at or below 0.013, and the remaining seven reach 0.03 to 0.13, with 10 dB noise at 0.80, yet recalibrating on same-attack negatives still leaves DA at 1.000 throughout (Appendix A.4), so the elevated rates

reflect a shift in the absolute score level rather than any loss of separability between the two classes. Detection also survives held-out re-synthesis and two-stage attacks (Appendices A.9 and A.11).

Imperceptibility. The host-bound carrier and attack-channel training cost some transparency, placing KeyBound at a wideband PESQ of 3.03, STOI of 0.981, and SNR of 24.6 dB. That is below WavMark (4.18) and Timbre (3.55) on PESQ, yet it is the highest intelligibility among the learned schemes, well clear of Timbre's STOI of 0.715 (Appendix A.7). The price is structural, since a psychoacoustic loss places the carrier where the host masks it, which is also where a codec or denoiser discards it, and sweeping the SNR budget moves denoiser bit accuracy from 0.998 to 0.792 as the mark grows quieter (Appendix A.10).

No-key decoding and false attribution. Without the key, KeyBound's detector recovers the payload at approximately chance, with a bit accuracy of 0.512 in the main pass and 0.499 in the per-clip pass of Appendix A.5, consistent with the single-sample behavior of Proposition 1, whereas all three baselines are keyless and yield their payload to anyone running the public decoder. For false attribution we lift a recovered carrier onto unrelated audio and ask whether it verifies, requiring the presence score to clear that method's own 1%-FPR threshold and the payload to decode. KeyBound verifies at 0.04 against 0.59 for AudioSeal and 0.227 for Timbre, though WavMark also resists (0.000), so host independence alone does not settle transplantability. KeyBound is nonetheless the only scheme evaluated supplying both properties at once, since each baseline either exposes its payload to any reader or accepts a transplanted carrier (Appendices A.6 and A.12).

Attack	WavMark	AudioSeal	Timbre	KeyBound
Clean	1.000	0.813	1.000	**1.000**
Signal distortion				
Noise (20 dB)	0.000	0.017	0.370	**0.997**
Noise (10 dB)	0.000	0.000	0.007	**0.980**
Lowpass	1.000	0.810	1.000	**1.000**
Highpass	1.000	0.807	1.000	**1.000**
Resample	0.947	0.727	1.000	**1.000**
Gain	0.993	0.760	1.000	**1.000**
Quantize	0.743	0.533	0.977	**1.000**
MP3	0.710	0.593	1.000	**1.000**
Smoothing	0.940	0.767	1.000	**1.000**
Echo	0.653	0.637	**1.000**	0.993
Re-synthesis				
EnCodec	0.000	0.010	0.003	**0.990**
Griffin-Lim	0.987	0.000	**1.000**	0.200
Denoiser	0.000	0.000	0.010	**0.973**

Table 2: True-accept rate of the attribution rule V_A across the full attack suite, which requires both conjuncts to hold.

Attribution under the verification rule. Table 2 reports V_A itself rather than its conjuncts. KeyBound attributes at least 0.97 of clips on clean audio and on twelve of the thirteen attacks, Griffin-Lim being the sole exception. The baselines attribute reliably under filtering and the milder edits, so the suite is not uniformly discriminative, yet they collapse under noise and re-synthesis, detecting or decoding but seldom both. The best any reaches there is 0.370 (Timbre at 20 dB noise), and none exceeds 0.010 under the codec or denoiser. AudioSeal never exceeds 0.813 even on clean audio, where a bit accuracy of 0.91 leaves one clip in five short of the budget. For KeyBound the presence conjunct holds at 1.000 throughout, so attribution is bound by the payload conjunct alone, and under Griffin-Lim a mean BA of 0.70 leaves only 0.20 of clips clearing 14 of 16 bits, where WavMark and Timbre pass nearly intact. We accordingly claim provenance *attribution* for KeyBound under signal distortion, the codec, and the denoiser, and *presence detection* alone under vocoder re-synthesis, where the exact payload no longer survives.

5.3 Ablation Studies

We remove each mechanism in isolation, all else fixed (Table 3). Every mechanism supplies one property, and the two that guard against an adversary cost a little imperceptibility.

The key mask. The key mask carries no ablation row, since XOR masking is dimension-preserving and leaves every other metric untouched, so the no-key decode already reported captures its effect in full (0.51 bit accuracy against 1.00 with the key). It is necessary but not sufficient, since the host anchor and attack-channel training supply the properties that a mask cannot.

	Transplant↓	EnCodec		Denoiser		PESQ↑
	(verify)	BA↑	DA↑	BA↑	DA↑	
KeyBound	**0.04**	**0.98**	**1.00**	**0.98**	**1.00**	3.03
without anchor	1.00	0.52	1.00	0.97	1.00	3.64
without EnCodec	0.05	0.54	0.92	0.98	1.00	3.45

Table 3: Ablation of KeyBound's mechanisms, removing the host anchor or EnCodec from the training channel.

The host anchor. Removing the anchor while keeping the key raises the transplant verify rate from 0.04 to 1.00, so every transplanted carrier verifies once the carrier has collapsed into one reusable pattern (cross-host similarity 1.00 against 0.53, Appendix A.12). The anchor is more than a security mechanism, as the host-independent variant also loses most of its payload under the codec (EnCodec BA 0.52 against 0.98) while presence detection stays at 1.00 in both, so conditioning on the host helps the payload survive re-synthesis rather than hindering it. Its only price is transparency, with a PESQ of 3.03 against 3.64 without the anchor.

Re-synthesis in training. Robustness to re-synthesis comes from the attack channel, not the architecture. Dropping EnCodec from that channel lowers EnCodec bit accuracy from 0.98 to 0.54 and costs presence detection too (DA 0.92 against 1.00), while the denoiser, which remains in the channel, is unaffected. The ablated model is also quieter at a PESQ of 3.45, since a mark never trained to survive re-synthesis need not carry the energy that survival demands.

Payload length. Capacity trades off against robustness. The 8-bit and 16-bit settings retain high bit accuracy on clean and denoised audio, whereas at 32 and 64 bits bit accuracy falls to 0.72 and 0.64 and does not recover with further training, which places reliable capacity at roughly 16 bits, a length at which the single-sample wrong-key bound already reaches $\beta \approx 2.1 \times 10^{-3}$ (Appendix A.13).

6 Discussion

Limitations. The principal open limitation is adaptive removal, which falls outside the wrong-key bound. The held-out remover R_{test}, with access to neither the detector nor the key, drives detection to 0.02 within twenty thousand steps at a PESQ of 4.42 (Appendix A.10), so our robustness claims cover fixed, non-adaptive transformations only, as for any post-hoc watermark whose attacker can train a remover. The clean-calibrated DA also realizes a higher false-positive rate under several attacks, up to 0.80 under 10 dB noise, although attack-matched thresholds leave DA at 1.00 (Appendix A.4). Key reuse is a further limitation. One key fixes one 16-bit mask, so a deployment that reuses a key is exposed to known-payload and multi-sample inference, which Proposition 1 does not cover and which we do not evaluate. Transplant resistance is likewise tested only against direct residual transplantation, and an adversary who re-embeds a lifted payload through an approximated embedder remains untested. Our evidence is 300 one-second clips of read English speech, and other speakers, languages, longer recordings, and long-term key reuse still await study.

Deployment scope. KeyBound is not always the right choice, since benign processing with no adversary is served more cheaply by a transparent mark. KeyBound suits settings where a party may read, remove, or fabricate an attribution, where keyless designs lack no-key protection and vary in transplant resistance, and where a verifier needs a keyed, host-bound rule with a conditional single-sample wrong-key acceptance bound (Proposition 2).

7 Conclusion

Distortion-only evaluation does not suffice for speech provenance, where an adversary may read, remove, or transfer a mark. KeyBound answers with a key mask, a host-bound carrier, and attack-channel training. On read speech it keeps detection under re-synthesis (spectral denoiser DA = 1.00, BA = 0.98), decodes at chance without the key, and rejects transplanted carriers, with detection transferring to held-out DAC and BigVGAN and to compound attacks as exact payload recovery degrades. The evidence covers one-second read English speech and fixed, non-adaptive attacks, and adaptive removal and key reuse remain open. We therefore read keyed, host-bound watermarking as a step toward treating provenance as the adversarial attribution problem it is, one in which keying, host binding, and attack-channel training each supply a distinct property.

AI USE STATEMENT

We used generative AI tools to polish the writing of this manuscript, improving its grammar, phrasing, and readability. The authors reviewed and verified every AI-assisted revision and take full responsibility for the content of the paper, including its scientific claims and every reported result.

ETHICS STATEMENT

This work aims to make synthetic speech attributable to its source, as a defense against impersonation, fraud, and fabricated spoken evidence. Our evaluation also studies attacks on watermarks, namely payload reading without the key, carrier transplantation onto unrelated audio, and a learned remover that strips the mark without audible damage. An adversary could reuse these attacks against deployed schemes, and we report them because a provenance mark whose weaknesses stay unmeasured invites misplaced trust. Learned removal of this kind is already documented in prior work (Li et al., 2025; Zhao et al., 2024), and our experiments only measure it against KeyBound.

All experiments use LibriSpeech (Panayotov et al., 2015), a public corpus of read audiobook speech released for research, together with the released code and checkpoints of the baseline methods. The study records no new speech, involves no human participants, and collects no personal data beyond what the public corpus already contains.

A watermark verdict is probabilistic evidence and can be wrong. The wrong-key acceptance bound of V_A is conditional and single-sample, the clean-calibrated presence detector realizes a higher false-positive rate under several attacks, and an adaptive remover defeats the mark, as our experiments and Discussion report. A KeyBound decision should therefore not be the sole evidence of origin in legal or forensic settings, and an absent mark does not show that a recording is genuine.

REPRODUCIBILITY STATEMENT

Section 3.1 states the threat model, and Section 4 specifies the full method, including the key mask, the host-conditioned carrier, the keyed decoder, the training objective, and the verification rule with its assumptions. Appendix A.1 proves both propositions. Section 5.1 describes the data split, baselines, attacks, and metrics, and every method is evaluated on the same seeded 300 test clips under the same payloads. Appendix A.2 lists the architecture, optimizer, training curriculum, loss weights, and computing platform, Appendix A.3 gives the configuration of each baseline, and Appendix A.8 gives the parameters of every attack. Appendices A.4 and A.5 report the false-positive control and the confidence intervals that support the main results.

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*, 47(4):1423–1443, 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.

Max Costa. Writing on dirty paper. *IEEE transactions on information theory*, 29(3):439–441, 1983.

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

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Sergei I Gel'fand and Mark S Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

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

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367–6371. IEEE, 2022.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in neural information processing systems*, 36:27980–27993, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

Sang Gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *11th International Conference on Learning Representations*, 2023.

Kexin Li, Xiao Hu, Ilya Grishchenko, and David Lie. Harmonicattack: An adaptive cross-domain audio watermark removal. *arXiv preprint arXiv:2511.21577*, 2025.

Pengcheng Li, Xulong Zhang, Jing Xiao, and Jianzong Wang. Ideaw: Robust neural audio watermarking with invertible dual-embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4500–4511, 2024.

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

Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, and Neil Z Gong. Audiomarkbench: Benchmarking robustness of audio watermarking. *Advances in Neural Information Processing Systems*, 37:52241–52265, 2024b.

Yixin Liu, Lie Lu, Jihui Jin, Lichao Sun, and Andrea Fanelli. Xattnmark: learning robust audio watermarking with cross-attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 38987–39015, 2025.

Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, 2003.

Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech. In *ICLR Workshop on GenAI Watermarking*, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210. IEEE, 2015.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pp. 749–752. IEEE, 2001.

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.

Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel. Latent watermarking of audio generative models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji. Silentcipher: Deep audio watermarking. In *Proc. Interspeech 2024*, pp. 2235–2239, 2024.

Wei Song, Zhenchang Xing, Liming Zhu, Yulei Sui, and Jingling Xue. Demark: A query-free black-box attack on deepfake watermarking defenses. *arXiv preprint arXiv:2601.16473*, 2026.

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.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. Seanet: A multi-modal speech enhancement network. In *Proc. Interspeech 2020*, pp. 1126–1130, 2020.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Yaoxun Xu, Jianwei Yu, Hangting Chen, Zhiyong Wu, Xixin Wu, Dong Yu, Rongzhi Gu, and Yi Luo. Wake: Watermarking audio with key enrichment. In *Proc. Interspeech 2025*, pp. 5093–5097, 2025.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6199–6203. IEEE, 2020.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. *Advances in Neural Information Processing Systems*, 37:8643–8672, 2024.

A Appendix

A.1 Proofs

Proof of Proposition 1. The map $\mu \mapsto m \oplus \mu$ is a bijection, so an unknown uniform mask renders $\tilde{m}$ uniform. Recovering m requires $\mu(\kappa)$, which only the verifier holds.

Corollary 1 (Fixed-target wrong-key agreement)**.** *Under the single-sample uniformity assumption of Proposition 1, a wrong-key decode matches any fixed target with the same probability as it matches a uniformly drawn target.*

Proof. Fix any target $m_A^\star$. By Proposition 1 the recovered bits are uniform on $\{0,1\}^n$ in the adversary's view, and a uniform distribution assigns the same probability to agreement with $m_A^\star$ as to agreement with a uniformly drawn target. □

Proof of Proposition 2. By Proposition 1 and Corollary 1, $\hat{m}(y, \kappa_A)$ is uniform on $\{0,1\}^n$ in the adversary's view, so $\Pr[d_H(\hat{m}, m_A^\star) \le \rho] = \sum_{i \le \rho} \binom{n}{i} 2^{-n} = \beta$, and the presence indicator, being an indicator, can only lower it.

A.2 Training Details

The generator and detector are SEANet-style convolutional networks (Tagliasacchi et al., 2020) with a 128-dimensional latent, the host analysis is a fixed log-mel front end with a small learned projection, and the payload is n=16 bits masked by a 16-bit key. We train with AdamW at a learning rate of 2×10^{-4} for 30,000 steps on batches of 12 one-second clips. A curriculum forms the code on a loud mark before tightening the SNR target from 6 to 22 dB over the first 7,000 steps, enables the learned remover after a 7,000-step warmup, ramps the false-attribution term over 3,000 steps and the perceptual terms over 5,000 after a 2,000-step delay, and raises the attack-channel probability from 0.25 to 1. The objective weights are λ_m=12, λ_d=1.5, λ_p=0.5, and λ_f=1, with the time–frequency masking and SNR-hinge terms at unit weight, and the remover is updated once per generator step at the same learning rate under a carrier scale of α=1.

All training and evaluation run on a single NVIDIA RTX 6000 Ada GPU (48 GB) hosted by an Intel Xeon w7-3565X with 502 GB of memory, under Ubuntu 22.04 LTS with CUDA 12.8, Python 3.10, PyTorch 2.8.0, and torchaudio 2.8.0. A full 30,000-step training run takes about 1.6 hours on this hardware, and a full evaluation pass over the attack suite takes under fifteen minutes per method.

A.3 Baseline Protocol

At submission time the official XAttnMark project page[1] offered the paper, results, and demonstrations but no public inference code or checkpoint, leaving us unable to run it under our attack protocol. WavMark is run on 1.1 second clips because its decoder partitions the signal into chunks of $16{,}000 + \lfloor 16{,}000 \times 0.1 \rfloor = 17{,}600$ samples and needs one complete chunk, which a 1.0 second clip cannot supply. Timbre runs at its native 22.05 kHz with a 10-bit payload and is attacked with the same algorithms at that rate. We keep every baseline at the configuration its authors released rather than retrain it to match KeyBound, so payload length, clip length, and sampling rate differ across methods, and the comparison is not fully controlled.

A.4 False-Positive Control

DA calibrates τ on clean unmarked audio, so we also score unmarked clips under each attack and report the realized false-positive rate at that same threshold (Table 4). Including the clean condition, seven of the fourteen evaluation conditions realize a rate at or below 0.013, so their DA means what it says. The remaining seven sit between 0.030 and 0.127, with EnCodec lowest and the denoiser highest, while 10 dB noise reaches 0.803, a point at which a reported DA of 1.000 is in fact realized at roughly an 80% false-positive rate rather than 1%. None of this costs detection, since the matched-threshold DA is 1.000 for every one of these attacks, which shift the absolute score level

[1] https://liuyixin-louis.github.io/xattnmark/

while preserving the margin between marked and unmarked audio. Under the denoiser the two distributions do not overlap at all, as the lowest-scoring marked clip still outscores the highest-scoring unmarked one, so enforcing a true 1% rate, or even 0.1%, leaves DA unchanged. Among the baselines only AudioSeal drifts comparably, and it does so more severely, as its EnCodec DA of 0.85 sits at a 0.58 false-positive rate and falls to 0.09 once matched, leaving most of that apparent detection a codec artifact rather than a recovered watermark. WavMark and Timbre hold the rate near zero, because their decode-derived scores cannot false-positive on unmarked audio at all.

A.5 Confidence Intervals

Table 5 reports 95% confidence intervals over the test clips, with exact Clopper–Pearson intervals for the binary rates and a 2000-resample clip-level bootstrap for the mean bit accuracies. They support the trends the aggregate tables report, with detection at 1.00 under codec and denoiser and no-key decoding at chance, and no KeyBound interval meets the matching AudioSeal one on detection, on denoiser bit accuracy, or on transplant verification. A dedicated per-clip pass with separately drawn keys produces them, so a point estimate may differ slightly from the aggregate tables wherever the key draw affects the quantity being measured.

Metric	Value	95% CI
Detection / security rates (Clopper–Pearson)		
KeyBound DA, denoiser	1.000	[0.988, 1.000]
KeyBound DA, EnCodec	1.000	[0.988, 1.000]
AudioSeal DA, EnCodec	0.850	[0.804, 0.888]
KeyBound transplant verify	0.043	[0.023, 0.073]
AudioSeal transplant verify	0.587	[0.529, 0.643]
Mean bit accuracy (clip-level bootstrap)		
KeyBound BA, denoiser	0.985	[0.980, 0.989]
KeyBound BA, EnCodec	0.983	[0.979, 0.986]
KeyBound no-key BA	0.499	[0.485, 0.513]
AudioSeal BA, denoiser	0.553	[0.538, 0.568]

Table 5: 95% confidence intervals.

A.6 No-Key Access and the Transplant Protocol

Table 6 collects the two adversarial axes on which KeyBound differs from the baselines. All three baselines are keyless, so their payload is readable by any party who runs the public decoder, and the no-key column records that as *public* rather than as a number. The transplant attack extracts the carrier residual $\delta = x_w - x$ from a marked clip together with its clean host, adds it unscaled to a different host, and tests verification. Because recovering δ requires the clean host, this is an oracle-residual attack stronger than the stated threat model, which grants the adversary only the marked recording, and KeyBound's 0.04 transplant verify rate therefore holds even under that attacker-favorable setting. Verification requires both conjuncts of V_A, and the presence conjunct uses each method's own 1%-FPR threshold rather than a shared constant, since presence scores across methods are on different scales. AudioSeal needs separate treatment, because its residual is not a low-energy carrier

Attack	DA (reported)	FPR	DA (matched)
Clean	1.000	0.007	1.000
Signal distortion			
Noise (20 dB)	1.000	0.047†	1.000
Noise (10 dB)	1.000	0.803†	1.000
Lowpass (4 kHz)	1.000	0.007	1.000
Highpass (500 Hz)	1.000	0.103†	1.000
Resample (0.8)	1.000	0.013	1.000
Gain ($\times 0.5$)	1.000	0.000	1.000
Quantize (8-bit)	1.000	0.007	1.000
MP3 (32 kbps)	1.000	0.007	1.000
Smoothing	1.000	0.003	1.000
Echo	1.000	0.050†	1.000
Re-synthesis			
EnCodec (6 kbps)	1.000	0.030†	1.000
Griffin-Lim	1.000	0.053†	1.000
Denoiser	1.000	0.127†	1.000

Table 4: False-positive control for KeyBound. Values marked † exceed twice the nominal rate.

Method	No-key accuracy↓	Transplant verify↓
WavMark	public	**0.000**
AudioSeal	public	0.587
Timbre	public	0.227
KeyBound (ours)	**0.512**	0.043

Table 6: No-key decoding and false attribution.

Method	PESQ↑	STOI↑	SNR (dB)↑
WavMark†	**4.18**	**0.993**	**37.4**
AudioSeal	3.44	0.968	18.2
Timbre	3.55	0.715	27.7
KeyBound (ours)	3.03	0.981	24.6

Table 7: Imperceptibility. † measured on 1.1 s clips.

Attack	Parameter / model	Strength
Noise	additive Gaussian	SNR 20/10 dB
Lowpass	biquad low-pass	cutoff 4 kHz
Highpass	biquad high-pass	cutoff 500 Hz
Resample	rational resample	factor 0.8
Gain	amplitude scale	$\times 0.5$
Quantize	bit-depth reduction	8-bit
MP3	LAME codec	32 kbps
Smoothing	moving average	window 7
Echo	delay + decay	0.1 s, 0.4
EnCodec	`encodec_24khz`	6 kbps
Griffin-Lim	STFT 1024/256	magnitude only
Denoiser	spectral gating	STFT 1024/256
Remover	SEANet U-net (ch 24)	marked→clean

Table 8: Attack configuration, read from the evaluation code.

at all (host-to-residual SNR 0.9 dB against 24.6 and 37.4 dB for KeyBound and WavMark), so a transplant carries over part of the source recording, and its rate should be read as an upper bound.

A.7 Perceptual Cost of the Mark

Table 7 reports what each mark costs on clean marked audio, scored against the unmarked host. WavMark is the most transparent scheme on all three measures, and the dagger records that it is evaluated on the 1.1 second clips its decoder requires rather than the 1.0 second clips the other methods use, matching the protocol of Table 1. KeyBound sits lowest on PESQ at 3.03 while holding the second-highest STOI at 0.981 and an SNR of 24.6 dB, so its carrier is paid for in perceived quality while intelligibility stays close to the best of the four. Timbre reverses that profile, reaching a PESQ of 3.55 on an STOI of 0.715, the lowest intelligibility here by a wide margin. Read alongside Table 1, this table gives the price side of the robustness reported there, and the operating-point sweep below traces how that price moves as the SNR budget is tightened or relaxed.

A.8 Attack Configuration

Table 8 lists the attack parameters read from the evaluation code. For KeyBound, WavMark, and AudioSeal, the attacks operate on 16 kHz inputs, except that EnCodec resamples internally to 24 kHz and back again across its round trip, while Timbre is attacked at its native 22.05 kHz. The two noise levels share one row, as do the two spectral operations built on the same 1024/256 STFT, namely the Griffin-Lim vocoder, which discards phase and re-synthesizes from magnitude alone, and the denoiser, which estimates a noise floor and gates against it. The learned remover in the last row is the held-out reconstruction remover R_{test} of the limitation study, which sees marked audio and clean targets and never sees the detector or the key.

A.9 Held-Out Re-Synthesis

EnCodec and the spectral denoiser appear in KeyBound's training attack channel, whereas DAC (Descript Audio Codec, Kumar et al., 2023, official 16 kHz checkpoint, all 12 quantizers) and BigVGAN (Lee et al., 2023, `nvidia/bigvgan_v2_24khz_100band_256x`) do not and therefore

Attack	DA	BA	PESQ	STOI	SI-SNR
DAC	0.92	0.70	2.99	0.96	10.5
BigVGAN	1.00	0.50	2.88	0.98	-2.7

Table 9: Held-out re-synthesis on the 300 test clips, with neither channel seen in training.

Budget	DA	BA	PESQ↑	STOI↑	SNR↑
1k	1.00	1.00	3.02	0.98	24.6
4k	0.17	0.66	4.18	0.99	32.8
10k	0.04	0.58	4.37	0.99	35.4
20k	0.02	0.55	4.42	0.99	36.9

Table 10: Held-out remover by training budget, with quality measured against the clean host.

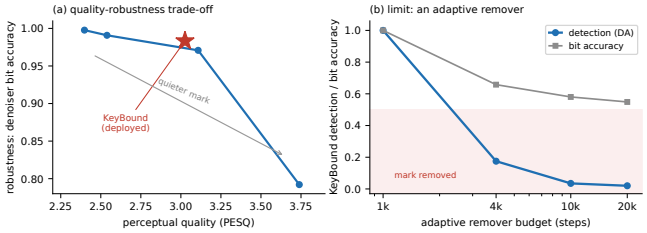

Figure 2: (a) Sweeping the SNR budget traces the quality–robustness frontier, and the deployed model (star) sits at a PESQ of 3.03 with 0.98 denoiser bit accuracy. (b) Detection and bit accuracy under a learned reconstruction remover (marked→clean, no detector access) of increasing training budget. The shaded region ($\mathrm{DA} < 0.5$) marks a removed watermark.

test attack generalization (Table 9). Presence detection transfers well to both, reaching DA 0.92 and 1.00 at the fixed LibriSpeech presence threshold, which we do not recalibrate and which agrees with the ROC TPR@1%FPR to rounding, while exact payload recovery degrades to BA 0.70 and 0.50, which reproduces on unseen channels the same two-level pattern that Griffin-Lim already shows in Table 1. The three quality columns compare the attacked output to the clean host, so they measure how much each channel damages the audio rather than how imperceptible the watermark is. SI-SNR is time-aligned by an integer cross-correlation lag, which matters for DAC, whose fixed 8-sample codec delay moves it from -15.0 raw to 10.5 aligned, whereas BigVGAN stays negative after alignment (-2.9 $\rightarrow$ -2.7) because mel re-synthesis rewrites the phase it cannot recover. Raw unaligned values are in the audit file.

A.10 Operating Point and Adaptive Removal

Figure 2(a) traces KeyBound's detection–quality trade-off as the SNR budget varies under a fixed 30,000-step schedule, where a quieter mark reaches higher PESQ while denoiser bit accuracy falls from 0.998 at the loudest setting to 0.792 at the quietest, and the deployed model sits between them at a PESQ of 3.03 with 0.98 denoiser bit accuracy. Figure 2(b) turns to the adaptive remover, against which detection survives a weak attacker (1.00 at one thousand steps) before falling to 0.17 by four thousand and to 0.02 by twenty thousand. Table 10 adds the quality of the removed audio over the same 200 clips, which is what makes this a genuine limitation rather than a nuisance, since once detection has collapsed at four thousand steps the remover output is closer to the clean host than the marked audio ever was (PESQ 4.18 to 4.42 against 3.03 for the mark itself), so removal costs the attacker nothing in audible quality and simply leaves the mark gone.

A.11 Compound Attacks

Chaining two transformations is a stronger threat than either applied alone. Table 11 reports detection accuracy at the clean-calibrated threshold under four such chains, on all of which KeyBound keeps DA = 1.00, while AudioSeal, already fragile under a single re-synthesis, drops to 0.41 and below. Its worst chain ends in vocoder re-synthesis and reaches 0.05, close to the 0.08 it scores under Griffin-Lim alone in Table 1, so the chain that finishes on its weakest single channel is the one that costs it the most.

Compound attack	KeyBound	AudioSeal
EnCodec → noise	1.00	0.41
Denoiser → resample	1.00	0.25
EnCodec → denoiser	1.00	0.22
MP3 → Griffin-Lim	1.00	0.05

Table 11: Detection accuracy under two-stage attacks.

Method	Carrier correlation↓	Cosine similarity↓
KeyBound	0.53	0.53
without host anchor	1.00	1.00

Table 12: Cross-host carrier similarity at a fixed payload and key, over 300 host clips.

Decoding setting	Bit accuracy
Correct key	1.00
No key / zero mask	0.51
Random wrong key	0.50 ± 0.04

Table 13: No-key and wrong-key decoding of KeyBound.

A.12 Host Binding and Wrong-Key Decoding

Table 12 checks host dependence directly. Holding payload and key fixed, KeyBound's carriers across 300 host clips show a mean pairwise correlation of 0.53, well below the variant without the host anchor, whose carriers are identical at correlation 1.00 and thus reduce to one reusable pattern. Cosine similarity gives the same ordering at the same two values, so the gap does not depend on which similarity we read. This is consistent with the transplant gap in Table 6, although transplant verification remains the direct false-attribution metric and these similarities support host dependence empirically rather than proving it. Table 13 reports keyed access over 300 marked clips, where the payload decodes at a bit accuracy of 1.00 to two decimals under the correct key and at chance under a zero mask or a random wrong key, the latter averaged over $K{=}10$ wrong keys per clip with the spread given as a standard deviation. A bit accuracy near 0.5 means the authorized decoder recovers no payload without the correct mask, which is the wrong-key behavior our no-key challenge measures and carries no claim about multi-message secrecy of the waveform itself.

A.13 Wrong-Key Acceptance and Payload Length

Proposition 2 gives a conditional single-sample bound on the probability that the attribution rule V_A accepts a decode made with the wrong key, and we restate here the constant and its scaling. Without κ_A the keyed decode $\hat{m}(y, \kappa_A)$ is uniform on $\{0,1\}^n$ in the adversary's view (Proposition 1, Corollary 1), so a false accept demands that a uniform draw fall within Hamming distance ρ of the registered payload $m_A^\star$,

$$\beta(n,\rho) = \Pr_{u\sim\mathcal{U}\{0,1\}^n}\left[d_H(u, m_A^\star) \le \rho\right] = 2^{-n}\sum_{i=0}^{\rho}\binom{n}{i}.$$

The presence gate $\mathbf{1}[s(y) \ge \tau_s]$ multiplies this by a value in $\{0,1\}$ and can only lower it, so the bound holds whatever the presence threshold and whatever attack inflates the presence-only false-positive rate. At the deployed operating point ($n = 16$, $\rho = 2$) this gives $\beta = 137/2^{16} \approx 2.1 \times 10^{-3}$, and because β falls combinatorially in n at fixed ρ (Table 14), the bound tightens by almost two orders of magnitude for every eight bits added. This is the statistical face of the capacity trade-off seen in the payload-length sweep, where an 8-bit payload is too short to bound anything meaningfully while lengths beyond 16 demand a payload robustness the carrier cannot sustain. The tolerance ρ moves the trade the other way, since a larger ρ raises the true-accept rate of V_A and enlarges β along with it, so we deploy $\rho = 2$ in every experiment reported here.

n (bits)	β
8	$1.4{\times}10^{-1}$
16	$2.1{\times}10^{-3}$
24	$1.8{\times}10^{-5}$
32	$1.2{\times}10^{-7}$

Table 14: Single-sample wrong-key acceptance bound $\beta(n, \rho{=}2)$ by payload length.